University of Alberta

Library Release Form

Name of Author: Sudeepa Bhattacharyya

Title of thesis: Structural proteomics: A high speed approach to identify function from structure

Degree: Master of Science

Year this degree granted: 2001

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as heretofore provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

University of Alberta

Structural Proteomics: A high speed approach to identify function from structure

by

Sudeepa Bhattacharyya ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Master of Science
in
Pharmaceutical Sciences

Faculty of
Pharmacy and Pharmaceutical Sciences

Edmonton, Alberta

Fall 2001

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled “Structural Proteomics: A high speed approach to identify protein functions from structure” by Sudeepa Bhattacharyya in partial fulfillment of the requirements for the degree of Master of Science in Pharmaceutical Sciences.

Abstract

As part of a "Structural Proteomics Initiative" at the Ontario Cancer Institute, we have solved the three dimensional solution structure of a protein with no previously known function or structure. The protein is a 77 residue peptide from the archaebacterium *Methanobacterium thermoautotrophicum ΔH* called *MtH*895. Using a combination of computational tools and modern NMR spectroscopy, we show (in chapter two of this thesis) that the protein adopts a well-defined fold similar to a class of redox proteins called thioredoxins/glutaredoxins. More detailed analysis shows that even though this protein has a glutaredoxin-like structure it appear to function as a thioredoxin. Indeed it appears to be the smallest known thioredoxin yet identified.

In chapter 3 of this thesis, the expression, purification, ^{15}N and ^{13}C/^{15}N double labeling of another protein, *MtH*807, from the same archeon, is described. Several optimization protocols have been discussed that resulted in a yield of ~40mg/L of protein from rich media and ~10mg/L from minimal media. More than 97% of the backbone amide resonances are visible in the ^{1}H-^{15}N-HSQC spectrum. ~25% of preliminary spin system assignments have been done.

Acknowledgements

I would like to express my deep gratitude to my supervisor, Dr. David S. Wishart, for his encouragement, support, guidance and help. I would also like to thank the committee members, Drs. Brian Sykes and Mavanur Suresh, for their valuable suggestions.

I have had the opportunity to work with a lot of my fellow graduate students over the last few years, and to them I owe my wholehearted thanks. In particular, I am grateful to the following colleagues: Haiyan Zhang (for help with computer-related issues), Hassan Monzavi (NMR spectrometer), Yamini Ramamoorthy (benchwork), and Ashnafi Abera (protein purification and isotopic labeling). I would also like to thank Alan Gibbs for introducing me to the software package, X-PLOR, and Bahram Habibi-Nazad for his help in carrying out the molecular dynamics simulations of *MtH*895 and T7 DNA polymerase interactions.

I am grateful to Drs. Adelinda Yee and Cheryl Arrowsmith of Ontario Cancer Institute for supplying me with labeled and unlabeled NMR samples of *MtH*895 and also the expression plasmid containing *MtH*807 DNA insert. I appreciate the help I received with NMR spectroscopy and the valuable instructions and guidance I got from Drs. Brian Sykes, Carolyn Slupsky, Leo Spyracopoulos and Stéphane Gagné over the past few years. As well, the financial support from Ontario Cancer Institute and PENCE is gratefully acknowledged.

Finally, I am extremely grateful to my husband, Abhijit, and my daughter, Trisha, for their endless patience, moral support and sacrifice over the last three years.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Amino acids:

A (Ala)	= L-alanine
C (Cys)	= L-cysteine
D (Asp)	= L-aspartic acid
E (Glu)	= L-glutanic acid
F (Phe)	= L-phenylalanine
G (Gly)	= L-glycine
H (His)	= L-histidine
I (Ile) .	= L-isoleucine
K (Lys)	= L-Iysine
L (Leu)	= L-Ieucine
M (Met)	= L-methionine
N (Asn)	= L-aspartic acid
P (Pro)	= L-proline
Q (Gin)	= L-glutamine
R (Arg)	= L-arginine
S (Ser)	= L-serine
T (Thr)	= L-theonine
V (Val)	= L-valine
W (Trp)	= L-tryptophan
Y (Tyr)	= L-tyrosine

2D	= two-dimensional
3D	= three-dimensional
Å	= angstrom
COSY	= correlation spectroscopy
CSI	= chemical shift index
DSS	= 2,2-dimethyl-2-silapentane-5-sulfonic acid
DTT	= dithiothreitol

E. coli	= *Esherichia coli*
EDTA	= ethylenediaminetetraacetic acid
FID	= free induction decay
Grx	= glutaredoxin
HMQC	= heteronuclear multiple quantum correlation spectroscopy
HSQC	= heteronuclear single-quantum coherence
INEPT	= insensitive nuclei enhanced polarization transfer
IPTG	= isopropyl □D thiogalactopyranoside
J	= coupling constant
kDa	= kilo Dalton
MtH	= *Methanobacterium thermoautotrophicum strain* ΔH
Mj	= *Methanococccus jannaschii*
NMR	= nuclear magnetic resonance
NOE	= nuclear Overhauser effect
NOESY	= nuclear Overhauser effect spectroscopy
OD_{600}	= optical density at 600 nm
PAGE	= polyacrylamide gel electrophoresis
PDB	= protein data bank
SDS	= sodium dodecyl sulphate
PBS	= phosphate-buffered saline
RMSD	= root mean square deviation
TOCSY	= total correlation spectroscopy
TSP	= 3-(trimethylsilyl)-propionate
Trx	= thioredoxin
UV	= ultraviolet

Chapter 1

Introduction

1.1 Structural proteomics and modern drug discovery

Now that many, ongoing, genome sequencing projects are coming to an end one could say that the genomic era is drawing to a close and the age of proteomics is about to begin. Simply stated, the ultimate aim of proteomics is to study all the gene products of all completed genomes. The enormous amount of sequence data generated by recent genome projects coupled with concomitant advances in molecular and structure biology techniques, have led to the concept of 'structural proteomics' or 'structural genomics'. Structural proteomics can be defined as the "*determination of three-dimensional protein structures on a genome-wide scale*" (*1*).

Knowing the structures of biological macromolecules is key to predicting, interpreting, modifying and identifying their functions and or active-site ligands. Gene or protein sequences, in most cases, reveal little about the exact mechanics of protein function or disease relevance. Hence, even though genomics has delivered a huge mass of raw information, it has largely failed to identify valid drug targets based on crude sequence information alone. In fact, it is estimated that >75% of all predicted eukaryotic proteins have no known cellular function *(2)*. Therefore, an important application of structural proteomics is to provide insight into a protein's function that is not otherwise detectable from sequence analysis alone. Often, proteins of similar function share structural homology in the complete absence of significant sequence homology *(3,4)*. As such, many newly sequenced proteins with no known structural or functional homologue may actually share unrecognized structural and functional

homology with known proteins. This 'hidden' homology can only be revealed by knowledge of their three-dimensional structure. On the basis of current estimates, Eisenstein *et al* (*5*) predicted that structural information will eventually provide functional clues for a large proportion of unannotated proteins.

A related application of structural proteomics is to help solve the so-called protein folding problem (*6,7,8*). The idea is to determine a sufficient number of three-dimensional structures to yield a complete representative set of protein folds (*1,2,4*). Most other structures could then be modeled from this basis set using computational homology modeling techniques. In other words, structural proteomics will put each protein within a comparative modeling distance of a known protein structure. It may also help with the development of structure-prediction algorithms that will eventually allow scientists to predict structure and function from sequence information alone (*10*).

As pointed out by Edwards *et al* (*10*), access to structural information on a genome-wide scale could have an immense impact on the pharmaceutical industry. Structural information can quickly reveal potential new drug targets, validate targets based on homology to other proteins or invalidate targets with structural properties that make them unsuitable for drug binding. An interesting example of how structural proteomics is having an important impact on the pharmaceutical industry can be seen with the work on the HIV genome. The Human Immunodeficiency Virus was completely sequenced in 1985 (*11*). Sequence analysis revealed the presence of around 15 proteins. Efforts have been ongoing since 1986 to solve as many structures in the HIV genome as possible. In the early 1990s an important breakthrough occurred when the

three dimensional structure of HIV protease was solved (*12,13*). It helped scientists design inhibitors that would interact with the protease in precise locations to specifically deactivate it. Structures for the HIV pol Nat and Tat proteins have also been completed (PDB id: 1AVV and 1TAC respectively). This research has helped to identify many new drug leads and cut down the number of AIDS deaths drastically in developed countries (*3*). Thus, structural proteomics can potentially help bridge the gap between genomics and drug discovery.

1.2 Determination of function from structure

Experimental approaches

Until recently, the function of a protein was usually ascertained before its structure was determined via X-ray or NMR techniques. However with the enormous amount of available sequence data and the recent advances in molecular and structure biological techniques, the current emphasis, in more and more laboratories, lies on the large-scale, high-throughput structure determination followed by functional assignment (*14,15,16*). An important characteristic of structural proteomics is that the structure determination projects are highly coordinated team efforts rather than the usual, distributed and uncoordinated efforts of individual structural biology laboratories. One of the primary goals of structural proteomics is to identify novel folds. This is accomplished by eliminating known folds from consideration, i.e., by excluding proteins that have clear homologues in the Protein Data Bank (PDB). The next step involves identifying those gene products that express and crystallize well for structure determination by X-ray crystallography. While X-ray crystallography is still the

primary tool for most structural proteomics efforts, NMR is particularly useful for identifying those proteins that express well but do not crystallize. However, candidate NMR proteins still need to remain soluble at concentrations needed for NMR spectroscopic studies (*17*). A key problem with all structural proteomics programs concerns the characterization of membrane proteins, which comprise roughly 20-30% of genomic sequences (*18*). These are normally excluded in structural proteomics efforts because the science of membrane protein structure determination is not yet advanced enough to be considered for high throughput techniques. Nevertheless, strategic target selection is an essential component of all structural proteomics efforts as this is key to maintaining the fast pace of structure determination and the rapid identification of novel protein folds.

Theoretical approaches

The most straightforward approach for predicting protein structure is to use the standard web-based bioinformatics or comparitive modeling tools such as Swiss-Model (*19*) and Modeller (*9*). Both of these Webservers have been adapted to perform high throughput automated structural modeling of entire genomes (*20*). Comparitive or homology modeling involves 3 steps: First one searches for sequence similarity to a member of a set of carefully selected sequences with known three-dimensional structure; second one uses the selected structure as a template to build a molecular model: and third one carefully validates the resulting models. An important limitation of the homology modeling process is that it requires a homologous protein of known structure. However, with new structures being solved at an increasing rate, a growing

proportion of genome sequences will likely have a homologous protein of known structure (*17*).

Beyond the sequence-based approach, which works best when sequence identity is, >35%, threading methods (*21,22*) can frequently identify sequences that have a similar fold but no apparent sequence similarity. Because threading can identify distantly related pairs of proteins, it can greatly increase the fraction of annotated proteins relative to traditional sequence-based methods (*23*). However, the key problem with current threading algorithms is that even though they can identify remotely related proteins, the corresponding structures differ considerably from the true structure and such inexact models are probably not sufficient enough to consistently provide clues to the protein's function (23). In addition to these "knowledge-based" methods there are more and more reports of ab-initio protein structure prediction methods (*24,25*) that have actually been able to identify novel or near-novel folds for small proteins (*26,27*). However the quality of these models are far from being accurate enough to replace existing experimental methods. These protein prediction methods can, however, assist in the target selection of the structural proteomics projects by identifying candidates with possible novel-folds (*17*).

Structure to function

Once the structure has been solved (either experimantally or theoretically), the most common way to extract information about it is to look for conserved sequence patterns within its structure (28). Although the structural context increases the

specificity of motif searching methods, it is not much more than an extension of standard sequence analysis methods in one-dimensional space.

A more vigorous approach would be to use three-dimensional motifs representing a specific functional site that can be systematically compared against the structure of interest. Thornton and colleagues have compiled one such library (http://www.biochem.ucl.ac.uk/bsm/PROCAT/PROCAT.html) in which a three-dimensional motif specifies the relative positions of certain atoms involved in a particular functional site. Both high-quality structures as well as low resolution ones obtained by the protein structure prediction methods can be scanned against a second structural motif library developed by Skolnick and co-workers in which the structural motifs are called ‘fuzzy functional forms’ (FFFs) (*29*). The resulting functional site analyses by these libraries have been claimed to be quite accurate (*17*). Overall, these *in silico* structure-based analyses can help to provide some clues concerning the general function of a protein of interest. These methods can also help in the detection of errors and can often augment any functional information provided by sequence analysis. Even though 3D motif searches cannot yet confidentially assign function at a detailed biochemical level, they can aid generate a hypothesis for biochemical function that could be readily tested experimentally in the laboratory. More often than not the structures actually provide an atomic level interpretation of existing functional data.

1.3 The Structural Proteomics Initiative at the Ontario Cancer Institute

With the aim to determine at least 10,000 different protein structures over the next five years, several coordinated structural proteomics pilot projects have been undertaken in the U.S, Canada, Japan and Germany (see Table 1.1, pg 27) with an aggregate public funding of more than U.S $100 million (*4,10*). A list of the pilot projects undertaken in structural proteomics all over the world, is presented in Table 1.2 (pg 27). These pilot projects which focus on relatively simple genomes are expected to deliver improved methods for high-throughput genomic-scale cloning, protein purification, sample preparation, structure determination, structure comparison and structure annotation. The methods developed by these efforts are expected to be applicable to more complex and more medically relevant proteomes such as those of pathogenic organisms or even humans.

One such initiative already underway in Canada is the ‘Structural Proteomics Initiative’ at the Ontario Cancer Institute under the supervision of Aled Edwards and Cheryl Arrowsmith (http://nmr.oci.utoronto.ca/arrowsmith/proteomics/). It aims at studying ~500 nonmembrane proteins from the proteome of the thermophilic archaebacterium *Methanobacterium thermoautotrophicum ΔH (MtH).* Their short-term goals are to assess the feasibility of such a high throughput effort, to estimate the percentage of proteins in a proteome that are immediately amenable to structural analysis and finally to evaluate the extent to which structures can provide insights into protein function.

Target Selection: MtH as a target organism

In the late 1970s, on the basis of rRNA phylogeny, Archaea (archaebacteria) was identified as a Superkingdom distinct from Bacteria (eubacteria) and Eucarya (Eukaryotes). Archaea display considerable morphological heterogeneity yet share a number of common biochemical and molecular-biological properties that distinguish them from "true" bacteria. Many archaeal species are adapted to extreme environments with respect to salinity, temperatures, alkalinity, acidity, etc. These extreme conditions have often posed a challenge in studying the molecular and mechanistic behaiviour of these unique organisms.

MtH is a thermophilic archaeon with an optimal growth temperature of 65°C. It belongs to a class of methane-producing archaebacteria that are obligatory anaerobes utilizing hydrogen, carbon dioxide and acetic acid to produce methane and carbon dioxide. It was originally isolated from the municipal waste-treatment facility in Champaign, IL, and has been widely studied as a representative of the methanogens that inhabit many biodegradation facilities (http://www.biosci.ohio-state.edu/~genomes/mthermo/). It was selected as a target organism for this structural proteomics project for several reasons. Firstly, it has a simple genome, yet shares many sequence and functional homologues with eukaryotes. Though they are phenotypically similar to eubacteria, archaebacteria are a distinct branch of the living world that often resembles eukaryotes more closely than prokaryotes, in their biology. For example, the archaeal transcription machinery contains a TATA-binding protein and a multisubunit RNA polymerase just like their human counterpart (*30*). Therefore, studies of the *MtH* proteins are expected to have some relevance to human proteins.

Secondly, the archaeal proteins are generally smaller and more compact than their eukaryotic homologues. For example, the archaeal homologue of the fifth largest subunit of RNA polymerase is one-third the size of its human counterpart (*31*). This renders them much more tractable for structural analysis. Thirdly, *MtH* proteins are stable at high temperatures. This facilitates their isolation and purification from *E. coli* to a great extent as the protein purification process is usually the rate-limiting step in most structural proteomics efforts. Infact, researchers at the OCI have shown that a single 70 °C heat-denaturation step (which precipitates most *E. coli* proteins) followed by a single chromatographic step to purify most recombinant proteins to the required level. This rapid, simple approach to protein purification is a critical achievement towards the success of any high-throughput effort. Last, but not the least, the *MtH* genome provides an excellent source of genes. The entire genome has been sequenced (30) and is freely available on the web. Furthermore, it was found that of all the thermophiles tested, the MtH genome was the most suitable for PCR-amplification.

Target protein selection

The *MtH* genome is comprised of 1,871 open reading frames (ORFs) of which ~30 % encode membrane-associated proteins. An additional 27% of known *MtH* proteins have clear structural homologues in the protein data bank (PDB). Of the remaining ~900 proteins, 424 were chosen as final targets for cloning, expression and subsequent structural studies. These represent an unbiased sampling of nonmembrane proteins of a single proteome of which 34% have a functional annotation, 54% are classified as 'conserved' (i.e., they are of unknown function and structure but have

bacterial homologues,) and 12% have no known structure or function. It has been hypothethized that these proteins may harbor new protein folds and/or reveal novel biochemical functions. It is also thought that they may provide clues to known functions through structural homology.

Implications

Since this project began (in 1999) several protein structures have been solved either by X-ray crystallography or by NMR spectroscopy. A list of the first 10 structures that have been solved (*1*) are shown in Table. 1.4 (pg 28) along with the functional clues extracted from the structural data.

Before their structures were solved these proteins had no known sequence homologues with any known three-dimensional structure. After their structures were determined, two things became obvious. First, no structure had a completely new fold as determined using the FSSP or SCOP classification databases. This suggetss, contrary to many expectations, that the discovery of entirely new folds will be a relatively rare event. Secondly, the structural similarity to known proteins yielded a range of information about the biochemical functions of these proteins, which were not evident from the sequence information alone. These could be readily tested by subsequent biochemical experiments in the laboratories.

There are some major limiting factors to the rapid progression of the project that are becoming more and more evident as the project continues. For example, limited access to expensive instruments such as NMR spectrometers and synchroton radiation sources has slowed the progress of several projects. In addition, poor expression levels

and unfavorable biophysical properties of a large number of proteins have also reduced overall throughput and productivity. However, in total, the project definitively demonstrates that structural proteomics is a feasible concept and the experimental techniques of X-ray crystallography and NMR spectroscopy can play a significant role in discerning the functions of proteins in a fast-track, genome-wide scale.

1.4 *MtH*895 and *MtH*807

As part of the structural proteomics initiative at the OCI, we selected two *MtH* proteins: *MtH*895 and *MtH*807. *MtH*895 is a 77 residue protein of previously unknown structure and function. However, it has atleast six conserved homologues both in the archaeal kingdom as well as in some thermophilic eubacteria and cyanobacteria. We have solved its three-dimensional structure using heteronuclear multidimensional NMR spectroscopy and have gained some valuable insights into its function from the structure using various computational tools. Our results indicate that even though it has a glutaredoxin fold it actually has thioredoxin-like activities. This is reported in detail in Chapter 2.

We have also expressed, purified, isotopically labeled and conducted preliminary NMR studies on *MtH*807. This protein has been identified as a putative thioredoxin in the *MtH* genome classification database. The molecular biology techniques involved in the expression, purification, ^{15}N and ^{13}C isotopic labeling of this protein are presented in detail in Chapter 3. Preliminary spin-system assignments based on both homonuclear and heteronuclear NMR spectra collected, are also presented in Chapter 3.

1.5 Thioredoxins and Glutaredoxins

Proteins catalyzing thiol-disulfide exchange reactions are required for many functions including electron and proton transport to essential enzymes like ribonucleotide reductase, for the formation of disulfides in protein folding and for general regulation of protein function by thiol redox control. Thioredoxins and glutaredoxins are small redox proteins that operate in thiol-disulfide reactions via two vicinal (CXXC) active site cysteine residues, which either form a disulfide (oxidized form) or a dithiol (reduced form) (*52*). The thioredoxin superfamily includes a growing number of proteins all having the same basic fold with a characteristic β/α/β/α/β/β/α arrangement of secondary stuctures and the active site located at the C-terminal end of a β-strand and followed by a α-helix (*52*). Interestingly, the thioredoxin fold has been identified in more than 150 different structures deposited in the PDB. Thioredoxin is reduced to the dithiol form by NADPH and thioredoxin reductase (together called the thioredoxin system). Glutaredoxin in contrast is reduced by the ubiquitous tripeptide glutathione (GSH) and GSSG in turn is reduced by NADPH and glutathione reductase (together called the glutaredoxin system). A large and growing number of functions in biological systems are known for the thioredoxin family of proteins. Besides redox regulation, thioredoxins have been implicated to participate in cell signalling, in controlling cellular growth and in regulating cell death or apoptosis (*53*). Thioredoxins also exhibit cytokine-like and chemokine-like activities. Both thioredoxins and glutaredoxins have been shown to play a key part in defending against oxidative stress (*53*).

1.6 NMR spectroscopy in solving protein structures

In recent years NMR has emerged as the single most powerful technique for determining the three-dimensional solution structures of proteins. Indeed, the recent and prospective advances in the techniques of NMR spectroscopy, such as partial deuteration, stronger magnets (up to 1GHz), superconducting probes, the TROSY experiments (*46*), triple resonance methods for resonance assignments (*47,48*), residual dipolar coupling analysis (*49*), automated and semi-automated methods (*50,51*) for resonance assignments and structure determination have significantly increased the speed and accuracy of structure determination. These advances will be critical to the success of the high-throughput structural proteomics projects (*4*). Even with the recent advances in X-ray crystallographic techniques, NMR will likely remain an indispensable tool due to its unique ability to rapidly and accurately obtain information about macromolecular dynamics and ligand binding properties. Steven Fesik and co-workers (*45*) have routinely used chemical shift perturbation to enumerate protein binding sites at a rate exceeding more than a thousand ligands per day. The technique known as SAR by NMR has been used to screen a library of 15,000 compounds, with about a mM sensitivity in the dissociation constant. Another significant contribution of NMR to structural proteomics may actually come from its potential to determine the structure of membrane proteins (*4*). However, the requirement of abundant quantities of isotopically labeled proteins and inherently poor expression levels of many membrane proteins will still be significant rate-limiting factors.

Physical principles of NMR

NMR spectroscopy deals with the interaction between the magnetic moments of atomic nuclei and a magnetic field (*32*). The magnetic moment of a nucleus (μ) is intimately connected to an intrinsic property called spin angular momentum ***I*** whose magnitude is quantized. According to quantum mechanics each subatomic particle (e.g., proton, neutron, electron), has a spin value of ½. Therefore, the overall spin property of an atom or nucleus results from a combination of the spins of its constituent subatomic particles. Some common nuclei, notably ^{12}C and ^{16}O, with even number of protons and neutrons, have $I = 0$, (i.e. no angular momentum), no magnetic moment and consequently are not NMR active. Isotopes with an odd number of neutrons and an even number of protons or vice-versa (e.g., ^{1}H, ^{15}N, etc) usually have a half-integral quantum number and generally produce excellent NMR spectra. Nuclei with odd numbers of protons and neutrons (e.g., ^{14}N) have more complex spin states and are less suitable for direct NMR observation. Fortunately, each of the four most abundant elements in biological molecules (H, C, N, & O) have one naturally occurring isotope with a non-zero nuclear spin and, therefore, are observable in an NMR experiment. While the naturally occurring isotope of hydrogen, ^{1}H is present at >99% abundance, other NMR active nuclei like ^{13}C and ^{15}N are present at much lower abundance (1.1% and 0.4%, respectively) and can be observed only after the target molecules are isotopically enriched. The ^{17}O isotope of oxygen, however, does not produce good NMR spectra as with a spin number $I = 5/2$, it has a non-spherical nuclear charge distribution giving rise to a quadrupole moment. This affects the relaxation time and consequently, significantly broadens the linewidth of the signal.

In the presence of an external magnetic field, the spin angular momentum of a nucleus with non-zero spin will cause that nucleus to undergo a cone-shaped rotational motion called 'precession' (Fig. 1.1a, pg 29). The rate of precession, referred to as Larmour frequency (ω_0) is dependent on the strength of the external magnetic field (B_0) and intrinsic properties of the nucleus reflected in its gyromagnetic ratio γ.

$$\omega_0 = -\gamma B_0 \qquad [1.1]$$

Each magnetic nucleus has $2I+1$ possible orientations and $2I+1$ corresponding energy levels with respect to an external magnetic field. For example, a spin-1/2 nucleus (^1H, ^{13}C) has two possible orientations, parallel and antiparallel, which corresponds to two different energy levels (Fig. 1.1b, pg 29). The energy difference between these two levels is directly proportional to the strength of the magnetic field .

$$\Delta E = \gamma h B_0/2\pi \qquad [1.2]$$

where h is Planck's constant. Note that in the absence of a magnetic field the spin angular momentum has no preferred directions and therefore leads to no detectable energy difference.

When placed in a magnetic field, a collection of magnetic nuclei, each absorbing a discrete amount of energy at its Larmour frequency, will partition themselves amongst the $2I+1$ available energy levels according to the Boltzmann distribution.

$$N_\beta/N_\alpha = e^{-\Delta E/kT} \qquad [1.3]$$

where N_β and N_α are the population of the lower and upper states respectively.

This population difference generates a net magnetization (M), which aligns with the external magnetic field (B_0) and remains in this equilibrium state. If a magnetic pulse is applied for a short period of time in such a way that it produces a second

magnetic field (B_1) perpendicular to the static field (B_0), it will drive M away from its equilibrium position by a so-called "flip angle". The magnitude of the flip angle depends on the time period and the field strength of B_1. A 90° pulse is therefore the time it takes for particular field strength to rotate the equilibrium magnetization 90° with respect to its equilibrium direction (which is the applied magnetic field usually set along the Z axis). Placing a metal coil in the XY-plane allows the recording of the oscillating current generated by the precessing magnetization. The precessing magnetization eventually falls back to equilibrium, with the XY magnetization slowly fading and the Z magnetization growing. This oscillating magnetic field can be detected by a coil, converted to an electric signal and recorded. This signal is called the free-induction decay (FID). The free induction decay can be converted from the FID (time-domain) to the frequency domain via Fourier transformation. The whole process is illustrated in Fig. 1.2 (pg 30).

Chemical shifts

The magnetic field at a nucleus is not exactly equal to the applied magnetic field. The electrons in a molecule surrounding a nucleus create a small magnetic field, which shields the nuclei slightly from the external field. The degree of shielding depends on whether a neighboring chemical group pushes or withdraws electron density from the nucleus through various inductive effects. Therefore, the Larmor frequencies of different nuclei vary due to their different chemical environment. This frequency change is called the chemical shift and it is an exquisitely sensitive indicator of chemical structure and composition.

The chemical shift (δ) is typically presented or measured in ppm (parts per million) and is defined as:

$$\delta = \frac{\text{Shift from standard (Hz)} \times 10^6}{\text{Spectrometer frequency (Hz)}} \text{ppm} \ . \qquad [1.4]$$

NMR spectroscopists prefer to measure chemical shifts in ppm instead of Hz because the former is independent of the magnetic field strength B_0. The chemical shift reference most commonly used is the signal of the methyl groups of tetramethylsilane (TMS), which, by definition resonates at 0 ppm. In protein NMR (2,2-dimethyl-2-silapentane-5-sulfonic acid) (DSS) is used equivalently. Different chemical groups have different chemical shifts and consequently chemical shift assignments provide a great deal of information for NMR spectroscopists. In proteins, for example, the signals of HN, H_α aromatic and aliphatic protons, all the backbone and side chain ^{13}C atoms and also ^{15}N signals in isotopically labelled proteins can easily be distinguished and frequently assigned on the basis of their chemical shifts. Additionally, the chemical shifts contain valuable information about protein secondary and tertiary structure.

Scalar Coupling

Nuclei, when close to one another, exert an effect on each other's effective magnetic field. This effect manifests itself in the NMR spectrum when the nuclei are experiencing different chemical environments or are chemically non-equivalent. If the separation between non-equivalent nuclei is less than or equal to three (sometimes

four) bond lengths, this effect is observable as a form of peak splitting or through the appearence of 'multiplets'. The seperation between the lines in a multiplet (in Hz) is given by the coupling constant J which is independent of the magnetic field strength. J couplings between pairs of protons separated by three covalent bonds (^3J, vicinal coupling) contains information about the intervening torsion angles and is described by the Karplus equation (*33*):

$$^3J = A \cos(\theta) + B \cos^2(\theta) + C \qquad [1.5]$$

where A, B and C are empirically derived constants and θ is the torsion (dihedral) angle.

The three-bond coupling constant between the intra-residue alpha and amide protons ($^3J_{HNH\alpha}$) is most useful for protein secondary and tertiary structure determination of as it can directly be related to the backbone dihedral angle ϕ.

Protein Structure Determination

Protein structure determination is a step-wise process initiated with the assignment of each resonance or NMR peak to a specific nucleus. The assignment strategies employed for any given protein strongly depends on whether the target protein is unlabeled or isotopically labeled. The archaebacterial protein, *MtH*895 whose solution structure is presented in this thesis, was available in both ^{15}N singly-labeled form as well as $^{15}N/^{13}C$ doubly labeled form. Therefore, both 3D ^{15}N-heteronuclear spectra and triple resonance spectra were used to complete the backbone and side-chain assignments for this protein.

Triple resonance experiments are frequently conducted on larger proteins (>10 kd) as the problems of spectral overlap can be markedly reduced. In triple resonance experiments the magnetization is efficiently transferred through ^{1}J or ^{2}J couplings (i.e. directly via the covalent chemical bonds), therefore, the transfer times are shorter and the signal losses due to relaxation are smaller than in homonuclear experiments. A prototype triple resonance experiment HNCA (*34,35*) is depicted in Fig. 1.3a (pg 31). Starting at an amide proton (H) the magnetization is transferred to the directly attached nitrogen atom (N) which is measured as the first spectral dimension. Then the magnetization is transferred to the C_α nucleus which is measured in the second dimension. Afterwards, the magnetization is transferred back the same way to the amide proton, which is measured in the third (direct) dimension. In each step magnetization is transferred via strong ^{1}J couplings between the different nuclei. The coupling which connects the nitrogen atom with the C_α carbon of the preceeding amino acid (^{2}J= 7 Hz) is only marginally smaller than the coupling to the directly attached C_α atom (^{1}J = 11 Hz). Thus, the nitrogen atom of a given amino acid is correlated with both its own C_α and the C_α of the preceeding amino acid. Therefore, in principle, it is possible to assign the protein backbone exclusively with an HNCA experiment, although in reality, more triple resonance experiments are needed to identify the cross signal of the preceding amino acid and to resolve degenerate resonance frequencies.

A very useful triple resonance experiment for performing side-chain assignments is known as the HCCH-TOCSY (*36*). In this experiment magnetization is transferred from a sidechain (or backbone) proton to the directly attached carbon atom, by ^{1}J

coupling to the neighboring carbon atoms and finally to their attached protons (Fig. 1.3b, pg 31).

Heteronuclear 3D spectra like ^{15}N-NOESY-HSQC (identifying through-space correlation) (*7*) and ^{15}N-TOCSY-HSQC (identifying through-bond connectivity) *(7)* can also be analyzed for assignment purposes. In these spectra, only the NOESY or TOCSY signals from those protons which are directly attached to a nitrogen are visible. This seperation makes 3D spectra significantly less crowded as compared to their 2D counterparts.

Structure calculation

After the sequential backbone and side-chain assignments of a protein are completed, the next step in the structure determination process is to estimate proton-proton distances from signal intensities derived from NOESY experiments. NOESY experiments measure the through-space correlation between nuclei that are close in space. This correlation, termed the Nuclear Overhauser Effect or NOE (*37*), is proportional to $1/r^6$, where r is the distance between two nuclei. The sign and intensity of the NOE also depends on the γ value of the interacting nuclei and the correlation time describing the motion of the interproton bond vector. As such, the NOEs fade quickly with distance and are not usually observed between protons that are more than 5 Å apart. However, the NOE is very sensitive to internuclear distances and therefore may be used to estimate proton-proton distances.

Secondary structural information can be reliably obtained by analyzing several parameters like short and medium range NOEs, $^3J_{HNH\alpha}$ coupling constants, amide

proton exchange rates and chemical shifts. In particular, a number of short and medium range inter-proton distances (<5 Å) are fairly unique to certain secondary structural elements. For example, alpha-helices are normally characterized by short distances between backbone amide protons (d_{NN}). They are also evident between beta protons of residue i and amide protons of residue i+1 ($d_{\beta N}$), as well as between the alpha proton of residue i and the amide protons of residues i+2, i+3, and i+4. Beta-strands are characterized by short sequential distances between the alpha proton and the adjacent amide proton ($d_{\alpha N}$). As well, the formation of beta-sheet results in readily observable NOEs between protons on adjacent strands (e.g., $d_{\alpha\alpha}$ and $d_{\alpha N}$) (*38*).

$^3J_{HNH\alpha}$ coupling constants also provide very useful information about secondary structure. For example, helical and extended conformations have very different values for phi (-60° and -120°, respectively) which result in measurable differences in $J_{HNH\alpha}$ coupling constants (*38*).

Hydrogen-bonded amide protons in regular secondary structures have measurably slower exchange rates with the solvent than amides in unstructured or flexible regions. Therefore, amide proton exchange rates can provide qualitative information about the location and stability of secondary structures. For example, continuous stretches of four or more slowly exchanging amide protons can indicate to the presence of a helix while alternating stretches of slowly and rapidly exchanging protons can indicate the existence of a beta-strand.

Chemical shifts can also contain useful structural information. H_α shifts as well as the Cα, Cβ and CO chemical shifts for all 20 natural amino acids have been shown to have a strong correlation with secondary structure (*39*). Based on these correlations

Wishart et al (*39,40*) described a simple method for secondary structure determination by analyzing the difference between residue specific backbone ^{1}H and ^{13}C shifts in proteins and that reported for the same residue in a "random coil" conformation. By combining these secondary constants with long-range NOE information one can usually determine a 3D solution structure.

For tertiary structure calculation, experimental distance restraints (NOEs) and torsion angle restraints (obtained from coupling constants) can be converted into a three dimensional structure by at least two different computational techniques: 1) distance geometry (*41,42*) and 2) simulated annealing (*43*).

The distance geometry method, in short, is based on a calculation of matrices of distance constraints for each pair of atoms from all available distance constraints, bond and torsion angles as well as van der Waals radii. This set of distances is then projected from the n-dimensional distance space into the three-dimensional Cartesian space, in which it determines the coordinates of all atoms of the proteins.

The simulated annealing process, on the other hand, is a molecular dynamics technique that takes place directly in Cartesian space. In this method, a starting structure is 'synthetically' heated to a high temperature (i.e. the atoms of the starting structure get a high thermal mobility) and then slowly cooled –much the same way that metal alloys are annealed. During the cooling phase the starting structure can evolve towards an energetically favorable, final structure under the influence of a force field derived from the constraints. Simulated annealing offers a more direct and consistent approach to finding a minimum energy structure through a very complex energy landscape or hypersurface.

Regardless of the chosen technique, what is essentially done is that a random coil starting structure is first generated from empirical data. The computer program then tries to fold the starting structure in such a way, that the experimentally determined constraints are as completely satisfied as possible. In order to achieve this, each experimentally measurable parameter is assigned a pseudo-energy potential, and standard energy minimization techniques are used to reduce the overall energy. Since the protein molecule can adopt a near infinite number of conformations, it is extremely important to identify as many experimental restraints as possible to restrict the conformational search space. Typically, 30-60 structures are calculated in order to sample the allowed conformational space. A detailed procedure using the simulated annealing protocol implemented in X-PLOR (*44*) to generate an NMR structure of *MtH*895 is described in Chapter 2.

1.7 References

1. Christendat , D. *et al*. Structural proteomics of an archaeon (2000) *Nature Structure Biology* 7(10): 903-908.

2. Brenner, S.E. & Levitt, M. Expectations from structural genomics. (2000) *Protein Sci.* 9, 197-200.

3. Service, R.F. (2000) *Science.* 287(5460):1954-1956.

4. Sali, A. (1998) Nature Structure Biology. 5(12), 1029-1032.

5. Eisenstein, E., et al. (2000), *Curr. Opin. Biotechnol.*, 11, 25–30.

6. Richards, F. M. (1991) *Scientific America*, 54-63.

7. Zhang, C. and DeLisi C. (1998) *J Mol Biol* 284: 1301-1305.

8. Wang, Z. X. (1998) *Protein Eng* 11:621-626.

9. Sali, A. and Blundell, T.L.(1993) *J.Mol.Biol.* 234, 779-815.

10. Edwards, A.M., Arrowsmith, C.H.& Bertrand des Pallieres (2000) *Modern Drug Discovery.* 3(7): 34-40.

11. Hahn, B.H., Gonda, M.A., Shaw, G.M., Popovic, M., Hoxie, J.A., Gallo, R.C., Wong-Staal (1985) F. *Proc. Natl. Acad Sci USA* 82(14):4813-4817.

12. Swain A.L., Miller, M. M., Green, J. Rich, D.H., Schneider, J., Kent, S.B. and Wlodawer, A. (1990) *Proc. Natl. Acad Sci USA* 87(22):8805-8809.

13. Harte W. E., Swaminathan S. Mansuri M. M., Martin, J.C., Rosenberg I.E. and Beveridge D.L. (1990) *Proc. Natl. Acad Sci USA* 87(220): 8864-8868.

14. Orengo, C.A., Todd, A.E. & Thornton, J.M. (1999) *Curr. Opin. Struct. Biol.* 9, 374–382.

15. Montelione, G.T. & Anderson, S. (1999) Structural genomics: keystone for a Human Proteome Project (news). Nat. Struct. Biol. 6,11–12.

16. Kim, S.H. Shining a light on structural genomics. (1998) *Nat. Struct. Biol. 5 Suppl*, 643–645.

17. Skolnick, J., Fetrow, J.S. & Kolinski, A. (2000) , *Nature Biotechnology* 18, 283-287.

18. Wallin, E. & Heijne, G.V. (1998) *Prot. Sci.* 7, 1029–1038.

19. Guex, N. and Peitsch, M.C. (1997), *Electrophoresis* 18, 2714-2723.

20. Sanchez, R. & Sali, A. (1998) *Proc. Natl. Acad. Sci.* USA 95, 13597–13602.

21. Rost, B., Schneider, R. & Sander, C. (1997) J. Mol. Biol. 270, 471– 480.

22. Jones, D.T. GenTHREADER (1999) *J. Mol. Biol.* 287, 797 –815.

23. Fischer, D. & Eisenberg, D. (1997) *Proc. Natl. Acad Sci USA* 94, 11929–11934 (1997).

24. Lee, J., Liwo, A., Ripoll, D.R., Pillardy, J. & Scheraga, H.A. (1999) *Proteins Suppl.* 3, 204–208.

25. Simons, K.T., Bonneau, R., Ruczinski, I. & Baker, D. (1999) *Proteins Suppl.* 3, 171–176.

26. Orengo, C., Bray, J.E., LoConte, L. & Sillitoe, I. (1999) *Proteins Suppl.* 3,149–170.

27. Murzin, A. (1999) *Proteins Suppl.* 3, 88–103.

28. Yu, L., White, J.V. & Smith, T.F. (1998) *Protein Sci.* 7, 2499–2510.

29. Fetrow, J.S. & Skolnick, J. (1998) *J. Mol. Biol.*281, 949–968.

30. Smith, D. R., Doucette-Stamm, L. A., Deloughery, C.et al. (1997) *J. Bacteriol.* 179, 7135-7155.

31. Mackereth, C. D., Arrowsmith, C. H., Edwards, A. M. and Mcintosh, L. P. (2000) *Proc. Natl. Acad Sci USA* 97: 6316-6321.

32. Kessler, H., Gehrke, M., Griesinger, C. (1988). Angewandte Chemie Internationalt Edition in English 27:490-536.

33. Karplus, M. (1959) *Journal of Physical Chemistry* 30:11-15.

34. Kay, L.E., Ikura, M., Tschudin, R., and Bax, A. (1990) *Journal of Magnetic Resonance* 89:496-514.

35. Grzesiek, S. & Bax., A. (1992) *Journal of Magnetic Resonance* 96:432-440.

36. Kay, L. E., Xu, G. Y., Singer, A. U., Muhandiram, D. R. & Forman-Kay, J.
J. Magn. Reson. Ser. B 101, 133-136.

37. Noggle, J.H. and Schirmer, R.E. (1971) The Nuclear Overhauser Effects, Academic, New York.

38. Wuthrich, K. NMR of Proteins and Nucleic Acids (1986) New York, John Wiley & Sons.

39. Wishart, D.S., Sykes, B.D. & Richards, F.M. (1992) *Biochemistry* 31:1647-1651.

40. Wishart, D. S. and Sykes, B. D. (1994) *J. Biomol NMR* 4:171-180.

41. Havel, T.F., and Wuthrich, K. (1985) *J. Mol. Biol.*182:281-294.

42. Williamson, M.P., Havel, T.F. and Wuthrich, K. (1985) *J.Mol Biol.* 182, 295-315.

43. Nilges, M., Clore, G.M. and Gronenborn, A.M. (1988) *FEBS Lett.* 229:317-324.

44. Brunger, A.T. X-PLOR. A (1992) System for X-ray Crystallography and NMR, Yale University Press, New Haven, CT.

45. Fesik, S.W. (1993) *J. Biomol NMR* 3 261-269.

46. Pervushin, K., et al. (1997). Proc. Nat. Acad. Sci. USA, 94, 12366-12369.

47. A. Bax, & S. Grzesiek (1993). *Acc. Chem. Res.* 26:131-138.

48. Zuiderweg, E.R.P. and Van Doren, S.R. (1994). *Trends Analyt. Chem.* 13, 24-36.

49. Prestegard, J.H., Tolman, J.R., Al-Hashimi, H.M. and Andrec, M.(1999). Protein structure and dynamics from field-induced residual dipolar couplings. In: Biological Magnetic Resonance: Structure Computation and Dynamics in Protein NMR (N.R. Krishna and L.J. Berliner, eds.)Vol. 17:311-355. Kluwer Academic-Plenum Publishers, New York.

50. Zimmerman, D.E. and Montelione, G.T.(1995) .*Curr. Opin. Struct. Bio.* **5,** 664-668.

51. Oschkinat, H. & Croft, D. (1994). *Meth. Enzymol.* 239, 308-318.

52. Holmgren, A. (1985) *Annu. Rev. Biochem.*, **54**, 237-271.

53. Arner, E.S. and Holmgren, A. (2000) *Eur. J. Biochem.*, 267, 6102-6109.

Table 1.1: Publicly funded academic projects.

Country	**Center**	**Funding (in millions)**
U.S.	NIH Structural Genome Centers	$50 (till 2000)*
Japan	NMR Park, Yokohama	$49
Germany	Protein Structure Factory, Berlin	$20
Canada	Univ. of Toronto	$23

*$150 million over next five years.

Table 1.2: List of pilot projects in structural proteomics

Organizers	**Organism**
Burley, S., Sali, A., Rockefeller University; Sussman, J., Weizmann Institute;	*S. cerevisiae*
Edwards, A., Arrowsmith, C., Ontario Cancer Institute & Univ of Toronto;	*M. thermoautotrophicum*
Eisenberg, D., U.C.Los Angeles; Terwilliger, T., National Struct Genomics Lab, Los Alamos;	*P. aerophilum*
Kim, S.-H., U.C. Berkeley;	*M. jannaschii*
Montelione, G., Rutgers University;	Metazoa
Moult, J., CABM;	*H. influenzae*
Yokoyama, S., Univ of Tokyo;	*T. Thermophilus, HB8*

Table 1.3: *M. thermoautotrophicum* genome statistics.

***MtH* genome encodes 1,855 open reading frames**		
46% with assigned, putative function.	**28%** with unknown structure & function but 'conserved' across archeal & bacterial kingdom.	**27%** with no known sequence or structural homologue.

Table 1.4. Functional insights derived from the first ten *MtH* protein structures that have been solved.

Protein	**Functional clue derived from structure**
MtH150—originally, a conserved protein of unknown function	Structurally similar to a nucleotidyl-transferase; contains an HXGH motif.
MtH152—also a conserved protein of unknown function	A flavin mononucleotide binding enzyme that needs Ni^{2+} ion for its catalytic activity.
MtH538—an unknown protein	Mg^{2+} binding enzyme; similar to AmiR-AmiC system
MtH129—unknown protein	An orotidine 5' monophosphate decarboxylase
MtH40—unknown protein	Similar to RPB10 subunit of RNA polymerase II; has a novel Zn binding motif.
MtH1048—unknown protein	Similar to the RPB5 subunit of RNA polymerase II
MtH1615—unknown protein	Involved in DNA binding or metabolism
MtH1699---identified as an archaebacterial translation elongation factor 1β from sequence analysis.	Structurally similar to human EF-1β or eubacterial EF-Ts but unlike them binds calcium.
MtH1184---unknown protein	Has a metal (not Zn) binding motif
MtH1175—unknown protein	Structurally similar to ribonuclease H superfamily

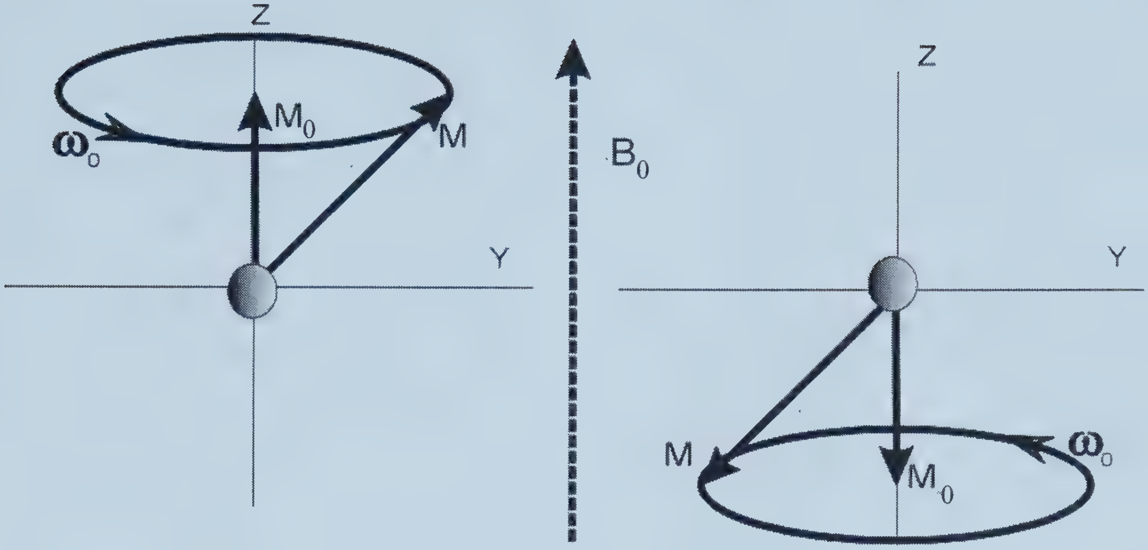


Fig. 1.1(a). The two allowed spin states for a nucleus of

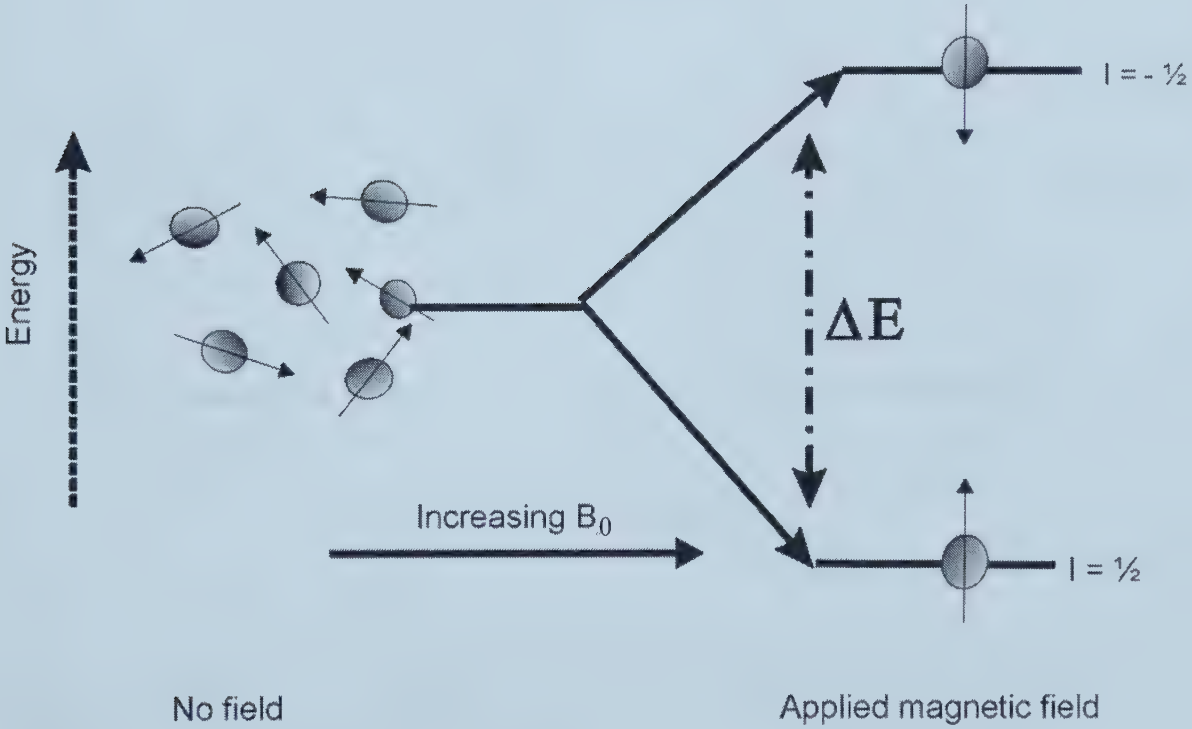


Fig. 1.1(b): Energy level diagram for spin I=1/2 in an applied magnetic field.

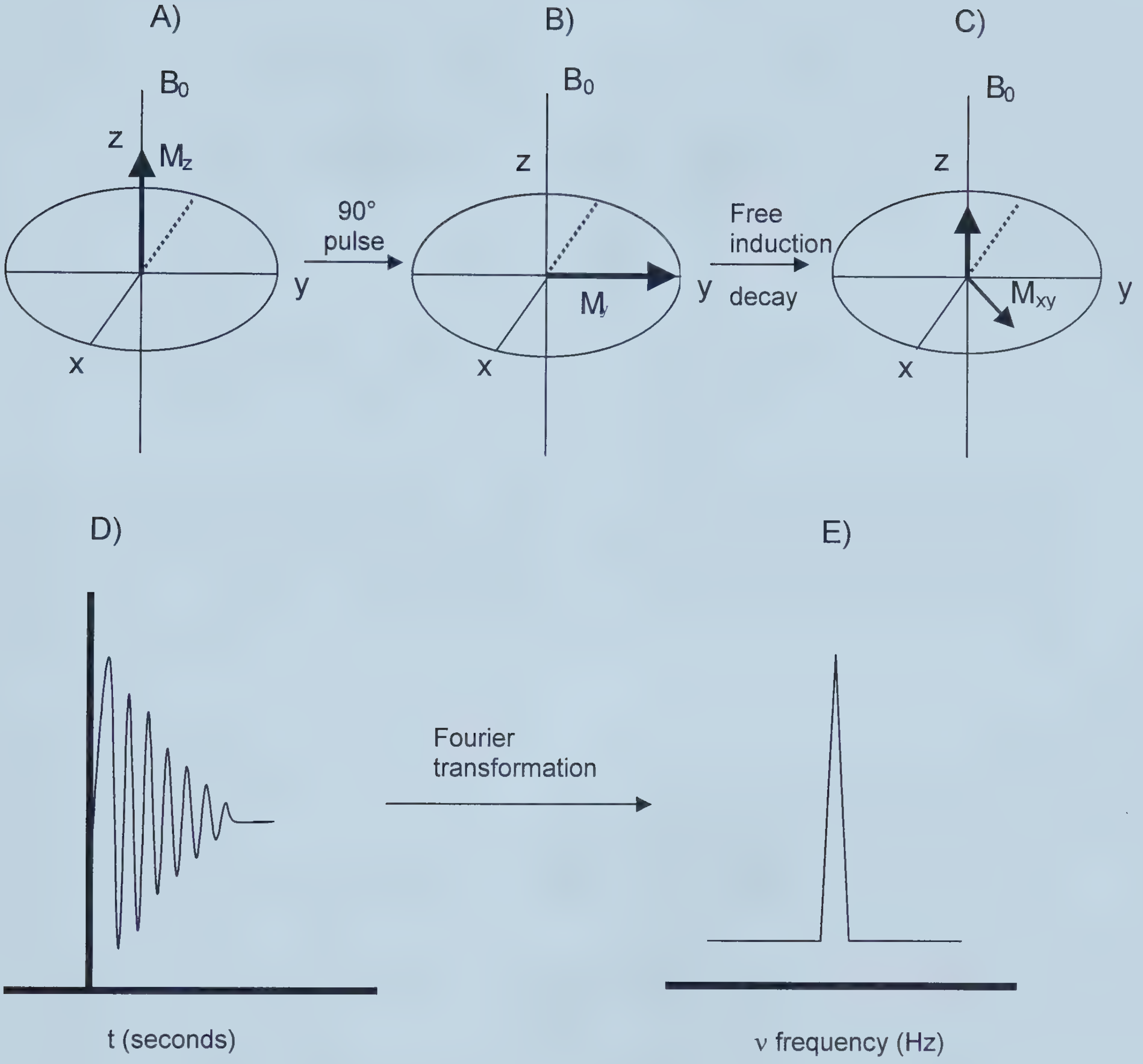


Fig. 1.2. Vector picture of pulse NMR. (a) Bulk magnetization M_Z at equilibrium in magnetic field $B_{0,}$ (b) M_Z rotates M_Y after 90° pulse, (c) M precesses at the Larmor frequency and returns back to equilibrium, (d) free induction decay (FID), (e) fourier transformation of FID yields the NMR spectrum.

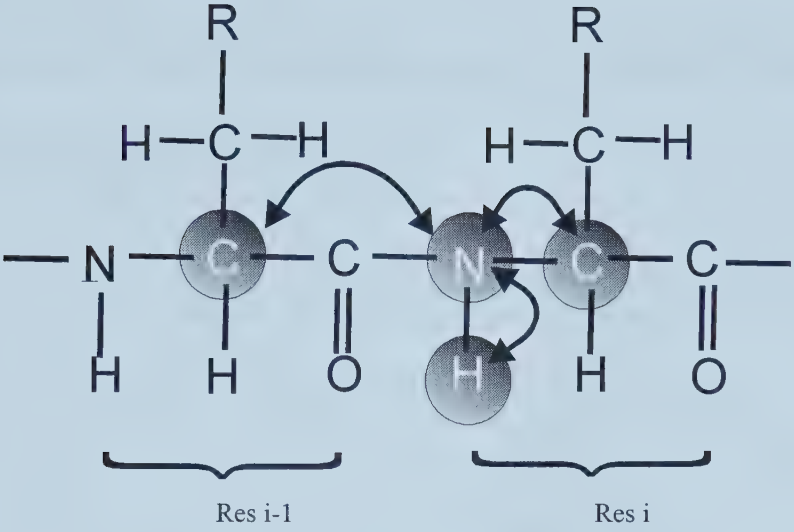


Fig.1.3 (a). Outline of magnetization transfer in HNCA experiment. Magnetization is first generated on the amide proton (H) and transferred to the attached ^{15}N via $^{1}J_{NH}$ coupling. Magnetization is then transferred to the intra as well as preceeding residue $^{13}C\alpha$ via $^{1}J_{NC\alpha}$ and $^{2}J_{NC\alpha}$ couplings respectively. Magnetizaation is then transferred back to the protons.

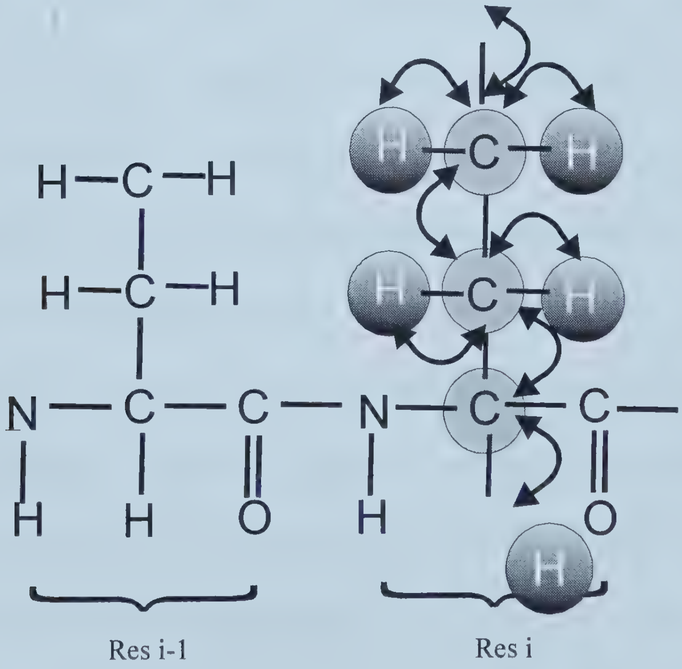


Fig. 1.3(b). Magnetization transfer in a HCCH-TOCSY experiment. Magnetization is transferred from a sidechain (or backbone) proton (darker circle) to the directly attached carbon atom (lighter circle), by ^{1}J coupling to the neighboring carbon atoms and finally to their attached protons.

Chapter 2

Identification of a Novel Archaebacterial Thioredoxin: Determination of Function through Structure

2.1 Introduction

The exponential growth in genome sequence data has placed increasing pressure on protein chemists to rapidly identify the function of many unknown or unclassified proteins. In cases where sequence comparisons fail to identify potential homologues or functional analogues, structural studies may go a long way towards revealing the function of the protein of interest (*1-4*). Because of the potential applications in functional classification, structural biologists are beginning to develop high throughput X-ray and NMR methods for rapid functional and structural characterization of proteins. Indeed, a number of international structural genomics initiatives are now underway aimed at solving the structure (and identifying the function) of a large number of proteins from a variety of model organisms (*4*). One such model organism is the archaebacterium *Methanobacterium thermoautotrophicum (ΔH)* -- a small thermopohllic bacterium first sequenced in 1996 (*5*). This particular archeon was chosen for this pilot project not only for its phylogenetic uniqueness, but also because it offered an opportunity to better understand the structural basis of the differential thermostability between thermophilic and mesophilic proteins (*1*). To date more than a dozen protein structures have been solved and characterized for this particular organism (*1*).

The genome of *M. thermoautotrophicum* (*MtH*) contains about 1870 proteins of which fewer than 50% have been assigned functions based on BLAST sequence

analysis (*6*). *MtH*895 is a small 77 residue protein identified as a conserved hypothetical protein with unassigned function. Because of its small size and good solution behavior, this protein was chosen for detailed structural analysis by NMR spectroscopy. Here we wish to report on the high-resolution structure of *MtH*895 and the subsequent identification of this protein (through sequential, structural and biochemical comparisons) as what appears to be the smallest known member of the thioredoxin family.

2.2 Materials and Methods

NMR sample preparation

Unlabeled, uniformly ^{15}N labeled and uniformly ^{13}C/^{15}N doubly labeled protein samples were generous gifts from the laboratory of Dr. Cheryl Arrowsmith (University of Toronto). The 234 bp *MtH*895 gene was expressed using the pET15b (Novagen) expression system in *E. coli* cells [BL21(DE3)] and purified by affinity chromatography using a Ni-NTA column as described elsewhere (*7*). The NMR samples were prepared by dissolving ~5 mg protein in 0.5 ml of 25 mM potassium phosphate buffer (pH 7) containing 200 mM NaCl and 10% D_2O. About 0.1 mM of 2,2-dimethyl-2-silapentane-5-sulfonic acid (DSS) was added to the samples as a chemical shift reference (*8*). The presence of 2 cysteines in the protein sequence suggested that thiol oxidation could lead to solubility problems. To keep the protein in a reduced state, excess (20 mM) dithiothreitol (DTT) was added to the NMR tube, which was then sealed with parafilm after degassing and refilling the tube with argon gas.

NMR spectroscopy

NMR experiments were recorded at 25°C on a Varian Unity 500 MHz spectrometer equipped with a 5 mm triple resonance probe and pulsed field gradient accessories. Two-dimensional ^{1}H-NOESY (with mixing times of 75 ms and 150 ms) (*9*), ^{1}H-^{15}N HSQC, 3D ^{1}H-^{15}N NOESY-HSQC (mixing time of 75 ms), ^{1}H-^{15}N TOCSY-HSQC (*10,11*) and HNHA (*12*) spectra were acquired with the unlabeled or ^{15}N labeled sample. In addition, ^{13}C/^{15}N-edited NOESY HSQC (mixing time of 75 ms) *(13*), HNCACB (*14*), CBCA(CO)NH (*15*), HCCH-TOCSY (*16*) as well as a 2D ^{1}H-^{13}C constant time HSQC experiments were collected on a ^{13}C-,^{15}N-doubly-labeled sample. Data were processed and analyzed using NMRPipe (*17*) and PIPP (obtained from Dr. Garret, NIH), respectively.

Assignments and Experimental Restraints

Beginning with the identification of ^{15}N and ^{1}HN chemical shifts from the ^{1}H-^{15}N HSQC spectra (Fig. 2.1, pg 60), spin systems were initially identified using ^{1}H-^{15}N TOCSY-HSQC data. Subsequently, backbone sequential assignments were completed using HNCACB and CBCA(CO)NH spectra using standard methods (*18*) and confirmed using ^{15}N/^{13}C NOESY HSQC data. Side chain assignments were completed using data from HCCH-TOCSY as well as 3D TOCSY and NOESY HSQC spectra. Stereospecific assignments of ^{1}Hβ protons were based on the intensity of ^{1}HN-^{1}Hβ cross-peaks in ^{15}N-edited- TOCSY and NOESY HSQC spectra (*19*). The methyl groups of Val and Leu were assigned stereospecifically based on the intensity of ^{1}HN- ^{1}Hγ, ^{1}Hα- ^{1}Hγ cross peaks, and the NOE intensity of the stereospecifically

assigned $^1H\beta$ protons to δ-methyl protons of Leu. NOE distance restraints were obtained using 3D ^{15}N-NOESY HSQC, 3D simultaneous $^{13}C/^{15}N$ NOESY HSQC as well as two-dimensional ^{1}H-NOESY spectra, all recorded with a τ_{mix} = 75 ms. The assigned NOE restraints were classified into four distance ranges: 1.8-2.8 Å, 1.8-3.5 Å, 1.8-5.0 Å, and 1.8-6.0 Å corresponding to strong, medium, weak and very weak NOE intensities, respectively. NOE peak intensities were measured by volume integration. Pseudo-atom corrections were added to the upper distance limits where appropriate (*20*). A 0.5 Å correction was applied to the upper bounds for NOEs involving methyl protons and nonstereospecifically assigned methylene protons. Torsion angle restraints were predicted using an in-house program (SHIFTOR – www.redpoll.pharmacy.ualberta.ca) based on the observed $^{13}C\alpha$, $^{13}C\beta$, $^1H\alpha$, and ^{1}HN chemical shifts and verified using the $^3J_{HNH\alpha}$ coupling constants obtained from a 3D HNHA experiment. A total of 41 φ backbone torsion angle restraints were used. These were assigned an uncertainty of ± 10° for residues in well-defined helical or beta-sheet regions. Backbone ψ dihedral angle restraints were derived using the same in-house routine as well as from an analysis of $d_{N\alpha}/d_{\alpha N}$ ratios (*21*) and assigned an uncertainty of ± 70°. Hydrogen bond restraints ($d_{O\text{-}HN}$ = 1.6-2.4 Å, $d_{O\text{-}N}$ = 2.6-3.4 Å) for slowly exchanging amide protons were identified from the pattern of sequential and interstrand NOEs involving ^{1}HN and $C_\alpha H$ protons and from the chemical shift index (*22*).

Structure calculations

Structures for *MtH*895 were calculated using X-PLOR 3.851 (*23*). Only NOE derived internuclear distance constraints were used in the first stage of structure generation. Initially a set of 50 structures was generated using the *ab initio* simulated annealing protocol applied to a template coordinate set. These structures were then regularized using the simulated annealing protocol of Nilges *et al.* (*24*). After this initial step, several ambiguous long-range NOE assignments were clarified by analyzing the resulting 50 structures. Subsequently the same simulated annealing protocol was repeated (three times) with additional and/or corrected NOE data. For the second stage of refinement, torsion angle restraints were introduced. The final set of 20 structures was selected on the basis that no interproton distance restraint violation could be greater than 0.5 Å and no torsion angle restraint violation could be greater than 5°. A total of 862 NOE-derived distance restraints [237 long range, 323 medium and short range (i - i+1, 2, 3 or 4) and 302 intra-residue], 102 dihedral angle restraints, and 46 hydrogen bond restraints were used to generate the final structural ensemble. The final set of 20 of structures was analyzed with PROCHECK-NMR (*25*) and VADAR (*26*). MOLMOL (*27*) was used to visualize all the structures and to calculate RMSD values.

Measurement of Thiol Ionization by Ultraviolet Absorbance

The thiol ionization constants for *MtH*895 were measured using the protocol of Dyson *et al* (*28,29*). The thiolate ion exhibits a stronger absorption at 240 nm (with an ε_{240} of about 4000 $M^{-1}cm^{-1}$) than the unionized thiol group. Therefore, the pK_a values

of the thiol groups can be monitored by UV absorption during pH titration. Specifically, 0.5 ml of a 1 mM *MtH*895 sample was reduced using 10X molar excess DTT. Excess DTT was removed by dialysis against the same buffer under an argon atmosphere. After dialysis, the *MtH*895 sample was diluted to a concentration of ~30 μM in a 0.1 mM EDTA, pH 6.0 and 100 mM potassium phosphate buffer. Thiol ionization was monitored by measuring the protein absorbance at 240 nm on a Pharmacia Biotech Ultraspec 3000 UV/Vis spectrophotometer. Small aliquots of 1 M NaOH or 2 M HCl were added to adjust the pH up or down. After each addition the change in protein concentration was noted at 280 nm. The concentration of *MtH*895 was calculated at 280 nm using an extinction coefficient of 1615 $M^{-1}cm^{-1}$ (*30*).

Modeling of Interaction between MtH895 and T7 DNA Polymerase

Molecular dynamics simulations of the binding interaction between *MtH*895 and T7 DNA polymerase were carried out using several approaches. Coordinates from the lowest energy conformer of *MtH*895 along with coordinates from the X-ray crystal structure of T7 DNA polymerase bound to *E.coli* thioredoxin (PDB entry: 1T7P) were used in these docking simulations. The probable binding sites and key residues involved in the ligand-enzyme interaction were identified from a detailed analysis of the 1T7P structure. Three proteins, *MtH*895, *E. coli* thioredoxin (PDB entry: 1THO) and *E. coli* Glutaredoxin (1GRX) were separately docked onto T7 DNA polymerase, both manually and automatically, using Swiss-PDB Viewer (*31*) and DOCKVISION (*32*) respectively. Different initial orientations of the ligands on T7 DNA polymerase were explored in order to obtain the best possible interactions of the

complexes, which were then further optimized to achieve the best geometry. The docked structures were solvated with an average of 2000 simple point charge (*33*) water molecules and subjected to steepest descent and conjugate gradient energy minimization using the GROMACS molecular simulation package (*35*). Initially, a 10 ps MD run was performed with positional restraints followed by a 1000 ps MD run without restraints. Weak coupling of the protein to a solvent bath of constant temperature (300 K) and constant pressure (1 bar) was maintained with a coupling time of 1.0 ps. For all energy minimization and MD simulations GROMOS-43B1 (*34*) and/or GROMACS (ffgmx) (*35*) force fields were used. The 20 lowest energy conformers were selected and energy minimized further using a conjugate gradient algorithm. All models were analyzed using WHAT IF (*36*) and PROCHECK (*37*). Quantitative comparisons between the binding interactions of the three proteins to T7 DNA polymerase were carried out using the Structural Thermodynamics Calculator (STC) (*38*).

2.3 Results

Solution Structure

Complete 1H, 13C and 15N chemical shifts for MtH895 are presented in Appendix A (BioMagResBank accession number 4991). Statistical parameters for the ensemble of 20 calculated structures are presented in Table 2.1 (pg 57). All structures (Fig. 2.2a, pg 61) exhibit good covalent geometry as indicated by low RMS deviations from idealized values and by low NOE, dihedral angle and van der Waals energies. For all 20 structures, ~99.6% of the main chain (ϕ, ψ) angles fall in the core or

allowed regions of the Ramachandran plot (Table 2.1, pg 57) as determined using PROCHECK-NMR and VADAR. With the exception of the active site loop (residues 9-13) and the C terminus (residues 75-77), all portions of the *MtH*895 structure are well-defined by NMR standards. The complete set of 20 *MtH*895 structures has been deposited with the Protein Data Bank (PDB accession: 1ILO).

*MtH*895 is composed of a four-stranded β-sheet sandwiched between two helices on one side and a third on the other (Fig. 2.2b, pg 61). The topological arrangement of the secondary structural elements and the overall fold clearly indicates that *MtH*895 has a glutaredoxin-like fold (*39*). The four β-strands include residues 2-7 (strand I) and 32-37 (strand II) which run parallel to each other and antiparallel to the β-hairpin formed by residues 53-56 (strand III) and residues 59-62 (strand IV). The first helix (helix I) is the longest helix and runs from residues 14-27. Helix II, which is on the same side of the central β-sheet as helix I, is composed of residues 40-46 while the third helix (helix III), connecting strands II and III, comprises residues 69-76. The axes of helices I and III run parallel to the β-sheet but in a direction opposite to the two parallel β-strands. The redox active site (Cys11-Ala12-Asn13-Cys14) is located just in front of the N-terminal side of the first helix. Like other thioredoxins and glutaredoxins (*39, 40*) this redox active site is solvent accessible on one side of the molecule but solvent inaccessible on the other. Close inspection of the active-site structure shows that the side chains of Ile6, Tyr7, Gln15, Met16, Leu17 and Val66 effectively block any approach to Cys14. *MtH*895 also has an exposed hydrophobic surface on the opposite side of its redox active site containing Cys11, Ala12, Asn13, Met40, Thr49, Ala50, Leu51, Pro52, Val66 and Ala67. This surface is similar to that

seen for other thioredoxins and glutaredoxins and is thought to have an important role in facilitating substrate binding and redox interactions (*41*,*42*). The relative positioning and general topology of this hydrophobic surface is presented in Fig. 2.3a (pg 62). Opposite to this hydrophobic face is a highly acidic surface (Fig. 2.3b, pg 62).

Structural comparison with other thioredoxins and glutaredoxins

A structural comparison was carried out between *MtH*895 and six other representatives of the thioredoxin/glutaredoxin superfamily: 1) T4 glutaredoxin, 2) *E. coli* thioredoxin (Trx), 3) *E. coli* glutaredoxin (Grx), 4) Human glutaredoxin 5) *Pyrococcus furiosus* protein disulfide oxidoreductase subunits N and C, and 6) thioredoxin-2 from *Anabaena sp.* This comparison revealed obvious similarities among all seven structures. Fig. 2.4 (pg 63) shows the structure-based sequence alignment, in which the secondary structural elements belonging to the common glutaredoxin/thioredoxin fold are numbered with Roman numerals. For all seven structures, the four β-strands forming the core β-sheet are well conserved. Overall, the secondary structures superimpose well with most insertions occurring at the N-termini or loop regions.

Structures similar to *MtH*895 were also identified by submitting a representative *MtH*895 structure to be searched against the Dali (*43, 44*), SCOP (*45*) and CE (*46*) databases. Table 2.2 (pg 58) lists the top hits from the CE database search. These same hits were also among the top hits in the Dali database search. The structures were assessed by their Z scores, number of matching residues and RMSD values. Interestingly, the top hits from all three databases belong to known

thioredoxins or to proteins with more than one thioredoxin domain. Notably glutaredoxins were largely absent from these lists.

Primary Sequence Analysis

After structural analysis by NMR revealed the existence of a glutaredoxin-like fold, we conducted a more detailed analysis using PSI BLAST (*6*). This database search clearly revealed that *MtH*895 belongs to the thioredoxin superfamily. After seven PSI BLAST iterations a list of approximately 150 proteins could be identified as *MtH*895 orthologs, with the vast majority being bacterial or archaebacterial thioredoxins/glutaredoxins. However, the level of sequence identity of *MtH*865 to any known member of the thioredoxin superfamily rarely exceeded 20%.

Interestingly, *MtH*895 exhibits a significant level of sequence identity (34-44%) to a group of previously unidentified proteins from several archaea including *Methanococcus jannaschii*, *Thermotoga maritima* and the cyanobacterium *Anabaena sp.* (Fig. 2.5, pg 64). It is worth noting that these primitive organisms already have previously identified thioredoxins/glutaredoxins, but the level of sequence identity of their known thioredoxins/glutaredoxins to any of the members of this new group of proteins, does not exceed 28%. For example, *MtH*895 has 28% sequence identity to the putative *Methanobacter* thioredoxin, *MtH*807. In *M. jannaschii*, the conserved hypothetical protein, *Mj*0581 has 42% sequence identity to *MtH*895, as opposed to 21% sequence identity to its recently characterized thioredoxin-like paralog *Mj*0307 (*47*).

Thiol ionization by UV spectroscopy

The stability of the active site disulfide bond is known to vary widely among the different members of the thioredoxin and glutaredoxin family (*48*). This variability has been shown to bear direct correlation with the pK_a value of the most accessible of the two cysteine thiols. The N-terminal cysteine of the active site –CXXC- motif is known to be solvent exposed and to have higher reactivity and a lower pK_a than the C-terminal cysteine. Normally, in glutaredoxins, the thiol pK_a values of the more exposed cysteine is less than 5 whereas in thioredoxins it is mostly above 6.5 (*48*). On the other hand, the pK_a value of the more buried cysteine is usually very high (above 9.0) in both thioredoxins and glutaredoxins (*49-53*). In an effort to determine whether *MtH*895 was more similar to a glutaredoxin or a thioredoxin, we determined the pK_a values of its two thiol groups using UV spectroscopy. On titration, the ε_{240} increased by a little over 8000 $M^{-1}cm^{-1}$ indicating both thiol groups were fully ionized (Fig. 2.6, pg 65). Interestingly, both thiol groups titrated at the same pH giving a pK_a of ~6.7 (as determined by the Henderson-Hasselbach equation). These pK_as are more typical of a thioredoxin than a glutaredoxin. This appears to add more evidence that *MtH*895 is more like a thioredoxin than a glutaredoxin.

Interaction between MtH895 and T7 DNA Polymerase

Glutaredoxins can be distinguished from thioredoxins on the basis of their interaction (or lack thereof) with T7 DNA polymerase. In order to assess if *MtH*895 could favorably interact with T7 DNA polymerase (pol) we carried out molecular docking studies between T7 DNA polymerase and *MtH*895, *E. coli* Trx (*54*) and *E.*

coli Grx. The binding interactions of each complex were quantitatively evaluated by calculating accessible surface area, thermodynamic parameters, as well as association and dissociation constants (K_a and K_d, respectively) using the STC program (*38*). The results are presented in Table 2.3 (pg 59). The calculated free energy change on binding between *MtH*895 and T7 DNA pol is comparable to that between *E. coli* Trx and T7 DNA pol. On the other hand the predicted binding energy between *E.coli* Grx and T7 DNA pol is approximately five times weaker. Similarly, the STC calculated K_a is only on the order of 10^3 for the T7 DNA pol-*E. coli* Grx interaction whereas it is on the order of 10^{17} and 10^{20} for the T7 DNA pol-*E. coli* Trx and T7 DNA pol-*MtH*895 interactions, respectively. This result shows that MtH895, on the basis of its predicted ability to bind with T7 DNA pol, resembles a thoiredoxin much more than a glutaredoxin. Fig. 2.7 (pg 66) shows the interactions that have been identified between *MtH*895 and T7 DNA pol from the majority of the conformers analyzed via our molecular dynamics and molecular docking simulations.

2.4 Discussion

Thioredoxin or Glutaredoxin?

The structure of *MtH*895 was solved as part of a pilot structural proteomics project initiated by the Ontario Cancer Institute *(1)*. It was aimed at determining the feasibility of using NMR to solve the structures of a large number of proteins and to assess the extent to which these structures could provide insights into their function. Before the structure of *MtH*895 was solved, this protein had no known sequence homologue with either an assigned function or a known three-dimensional structure.

After the solution structure was determined, it became obvious that *MtH*895 exhibited a glutaredoxin-like fold. However, the question quickly arose is it really a glutaredoxin? Glutaredoxins are small, ubiquitous proteins that are important for redox regulation of protein function and signaling. They accept protons from glutathione and transfer those protons to various protein substrates. Our suspicions were raised about the true function of this protein when a literature search revealed that *Methanobacterium thermoautorophicum* does not contain glutathioine or a glutathione-like cystolic thiol (*42*). Glutathione is essential for glutaredoxin function. Furthermore, all previous studies of putative archebacterial glutaredoxins concluded that these putative glutaredoxins functioned essentially as thioredoxins rather than glutaredoxins. That is, they exhibited no activity as glutathione mixed disulfide reductants (*42, 47*). To confirm our suspicions we undertook a detailed investigation to determine the true nature of *MtH*895 through 1) sequence comparison; 2) structural comparison; 3) structural analysis and 4) biochemical analysis.

Sequence comparison

As indicated previously, there is little similarity in the primary sequences between *MtH*895 and the members of the thioredoxin/glutaredoxin superfamily. Among those exhibiting the highest level of sequence identity was a group of previously unidentified or putative archebaterial thioredoxins. Sequence comparisons between *MtH*895 and "standard" thioredoxins indicate that only a few residues are fully conserved. These include the two active site cysteines, a cis-proline at the loop preceding β-III, a glycine residue between β-III and β-IV and a second glycine

immediately following β-IV. As a rule, the active site sequence -Trp-Cys-Gly-Pro-Cys- is highly conserved in thioredoxins whereas glutaredoxins have the conserved -Cys-Pro-Tyr-Cys in all but two members (T4-glutaredoxin has a Val in place of Pro and pig glutaredoxin has a Phe in place of Tyr). *MtH*895, however, has a unique -Cys-Ala-Asn-Cys- motif which is not found in any known member of the thioredoxin or glutaredoxin superfamily. We believe the absence of a Pro residue in the *MtH*895 active site may confer some conformational flexibility and therefore may have some effect on the stability of the redox reaction process in this enzyme. This is supported by some recent work by Ren *et al.* (*55*) who have looked at the crystal structure and thermal B factors of an archael disulfide oxidoreductase with an unusual active site motif. Interestingly, all members of this putative group of archaebacterial thioredoxins have unique -CXXC- motifs distinctly different from either thioredoxin or glutaredoxin active site motifs (Fig. 2.5, pg 64). Evidently, archaebacteria appear to have relaxed their requirement for an absolutely conserved sequence -Trp-Cys-Gly-Pro-Cys- active site for redox function.

Structure comparison

While direct sequence comparisons were not generally able to distinguish between the two possible functions for *MtH*895, structural comparisons were much clearer. As a general rule, significant differences exist between thioredoxins and glutaredoxins at their N-termini where the former typically has additional 15-20 residues. These extra residues form well-defined secondary structures that lay on either side of the central β-sheet core. However, these extra residues have not yet

been implicated as having any major role in the redox functionality of these proteins. *MtH*895 does not contain these additional N-terminal secondary structural elements, suggesting it is superficially more like a glutaredoxin. However when the structure of *MtH*895 was submitted to the Dali, SCOP and CE fold identification servers the top hits for all three database comparisons were members of the thioredoxin family – not the glutaredoxin family. These results strongly suggested that even though *MtH*895 has a glutaredoxin fold, it is more similar to the thioredoxins in structure.

Structure analysis

A more detailed structural analysis of the active site region of *MtH*895 revealed some important functional information as well. Most glutaredoxins have a highly conserved positively charged residue (Lys or Arg at position 8) in their protein-protein interaction site followed by a positively charged residue on the C-terminal side of their –CXXC- active site (*40*). An aspartic acid at the N-terminus of the third helix is also found in most glutaredoxins. These charged residues, which all lay near the active site, have been implicated in the participation of specific ionic interactions with glutathione (*56*). In contrast, most thioredoxins have neutral or hydrophobic residues in these sites. This difference arises because the interacting substrate for thioredoxins is a protein (thioredoxin reductase) rather than a small molecule like glutathione (for glutaredoxins). Interestingly, *MtH*895 does not have any of the characteristic charged residues of glutaredoxins, but rather, it has mostly neutral or hydrophobic residues in these sites. This strongly suggests that *MtH*895 cannot possibly interact with glutathione and therefore, by elmination, it must function as a thioredoxin.

Thiol Ionization

Our measurements indicate that *MtH*895's active site thiols have a pK_a around 6.7 which is more similar to the pK_a value of the exposed cysteine in thioredoxins than glutaredoxins (*57, 53*). However, the pK_a value of the buried cysteine is much higher in both these classes of proteins, than in *MtH*895. Interestingly, two other thioredoxin-like proteins appear to exhibit similar thiol pK_a values to those observed for *MtH*895. A thioredoxin-like protein (*Mj*0307) from a related archaeon, *Methanococcus jannaschii*, has a single pK_a value of about 6.28 for both its thiols (*47*). In addition, an *E. coli* thioredoxin D26A mutant shows simultaneous titration of both its thiols at pH 7.5 (*58*). In wild type *E. coli* thioredoxin, the N and C-terminal actie site cysteines have pK_a values of 7.1 and 7.9 respectively (*59*). The partial positive charge induced by a helix dipole from helix-II has been implicated in lowering the pK_a value of the N-terminal active-site cysteine in *E. coli* Trx. In the case of *MtH*895, the active site cysteines are in the equivalent secondary structural elements as in *E. coli* Trx and therefore would be expected to experience a similar helix dipole effect. However, in *E.coli* Trx the presence of Asp26 has been suggested to play a significant role in lowering the pK_a value of the N-terminal active-site cysteine compared to the other cysteine (*58*). *MtH*895 differs from *E. coli* Trx because it lacks a negatively charged residue at the equivalent position that could influence the N-terminal cysteine. The lack of a nearby negatively charged residue has also been suggested to be the reason for the simultaneous titration of the active site thiols in *M. jannaschii* thioredoxin (*47*). Other factors, such as the two non-standard amino acids between the *MtH*895 active

site cysteines may also influence their pK_a values. For example, Grauschopf *et al.* (*60*) have shown that mutation of the active site dipeptide Pro-His to Pro-Pro in DsbA (which is also a member of the thioredoxin superfamily) significantly changed the pK_a value of Cys30. Wang (61) suggested the scarcity of positive electrostatic potential surrounding the active site thiols may be the reason for the unusually high pK_a of Cys14 in T4 glutaredoxin as compared to the other glutaredoxins. Indeed, *MtH*895 also has a very few positively charged residues in and around the active site (Fig. 2.4, pg 63) which may be the reason for its relatively high pK_a value.

Interaction with T7 DNA polymerase

A key distinguishing feature between thioredoxins and glutaredoxins is their differential ability to bind T7 DNA polymerase. Most prokaryotic thioredoxins are known to interact with T7 DNA polymerase including *E. coli* Trx (54) and thioredoxins from *M. jannaschii* (47), *Thiobacillus ferroxidans* (62), *Corynebacterium nephridii* (63) and *Rhodobacter sphaeroides* (64). On the other hand, neither *E. coli* glutaredoxin nor T4 glutaredoxin (65) support T7 phage growth as glutaredoxins appear to be incapable of binding T7 DNA polymerase.

Fig. 2.7 (pg 66) shows the interactions that have been predicted to occur between *MtH*895 and T7 DNA pol. Both the location and the predicted interactions are very similar to those seen in T7 DNA pol bound to *E. coli* Trx (*54*). In both *MtH*895 and *E. coli* Trx, the site of interaction is located on the surface near the vicinity of the active site. Several amino acid residues in the loops containing the –CXXC- motif and those between βII and αII, αII and βIII, and βIV and αIII (using the

numbering scheme presented in Fig. 2.4, pg 63) are involved in binding. The CO group of Pro325 in T7 DNA pol is predicted to form a hydrogen bond with the NH of Val66 in *MtH*895, similar to the hydrogen bond it forms with Ala93 NH group located in an equivalent position in *E. coli* Trx. The Val329 NH group of T7 DNA pol forms a hydrogen bond with the Arg73 CO group in *E. coli* Trx. A similar hydrogen bond is also predicted to form between the Val329 NH group and Thr49 CO of *MtH*895, whose sequential position is identical to that of Arg73 in *E. coli* Trx. The second helix of *MtH*895 is predicted to participate in several important interactions with T7 DNA pol, similar to those seen in the *E. coli* protein. The active site cysteines in both proteins are predicted to form several inter and intra-molecular hydrogen bonds. Interestingly, N-terminal active-site cysteines in both complexes face outside and are more accessible compared to the C-terminal cysteines which seem to be rigidly trapped in a cavity. Nearly 65% of the residues in T7 DNA pol that are predicted to interact with *MtH*895 can be seen to interact with *E. coli* Trx in the X-ray crystal structure (*54*). The small difference seen in the interactions is probably due to the flexibility of the extended loop between helices H and H1 of T7 DNA polymerase. Important charge-charge interactions predicted in a majority of the conformers involve Arg65 (of *MtH*895) interacting with Glu272, Tyr326, Lys290, and Tyr320 of T7 polymerase. Similar interactions are also predicted between Lys61 and Asn13 of *MtH*895 with Glu259, Trp264, Glu330 and Tyr286, Thr327, respectively, of the polymerase. The corresponding residues, Lys90 and Pro34 in *E. coli* Trx, participate in similar interactions with the T7 DNA polymerase. This high degree of structural

equivalence along with the very favorable binding energies calculated via STC strongly suggest that *MtH895* functions as a thioredoxin rather than a glutaredoxin.

*Thermal Stability of MtH*895

One of the reasons for working with proteins isolated from *M. thermoautotrophicum* was to attempt to determine the underlying causes of thermostability in proteins. A number of structural and sequence-specific features have previously been suggested to confer high thermal stability to proteins in thermophilic organisms. For instance, the presence of an increased number of surface ion pairs, an increased number of hydrogen bonds, a higher proportion of aliphatic amino acids, and a larger polar surface area (for increased hydrogen bonding density with water) have all been suggested as underlying reasons for increased thermal stability in proteins (*66*). In *MtH*895, we have identified four surface ion pairs (according to the definition of Vogt (*67*)): Lys3 to Glu33, Lys3 to Asp57, Lys26 to Asp31 and Lys74 to Glu70. However, it is unclear whether these are in any higher proportion than what might be expected for a mesophilic protein of similar size. Assessments of the amino acid composition of *MtH*895 suggest that it has a higher than average proportion of aliphatic residues (Ala, Met, Ile. Leu) relative to most mesophilic proteins. The number of hydrogen bonds and proportion of polar surface area for *MtH*895 also appears to be higher than normal. Recently, Thompson and coworkers (*68*) noted that thermophilic proteins are generally smaller in size and more compact in structure. They are also more likely than their mesophilic counterparts to have deletions in exposed loop regions. Given that most thioredoxins and thioredoxin

domains are all more than 100 residues in length, *MtH*895, at just 77 residues is by far the smallest known thioredoxin. Most of the deletions in *MtH*895 appear to occur at the N-terminus and within its loop regions (Fig. 2.4, pg 63). Overall, it appears that many factors incrementally contribute to the thermostability of *MtH*895. However, given its unusually short length and compact size, we suspect structural compactness may be the most important contributor to the high thermal stability of this particular protein.

2.5 Conclusions

This project was undertaken to address a key question in structural proteomcs. That is: can we determine the function of a protein through its structure? In attempting to answer this question we deliberately restricted ourselves to using only the *MtH*895 sequence, its 3D structure and readily available computational or database tools (PubMed, BLAST, PSI-BLAST, DOCKVISION, CE, SCOP, CATH, etc.). We resorted to using biochemical methods (thiol titration) to confirm our conclusions only after it had become abundantly clear through structural and sequential analysis what the true function of this protein was. The thiol titration results essentially confirmed and helped rationalize what we had predicted through earlier structural analysis. While we were fortunate to have worked with a protein that had at least some resemblance to a known family of proteins, it is still important to remember that this resemblance was only ascertained after the structure was determined – not before. Furthermore, even though a functional assignment was made, we still do not know the exact metabolic role of *MtH*895 in *Methanobacterium thermoautrophicum*. Despite

this caveat, we believe this work reaffirms the potential for structural proteomics to be used in protein functional assignment. However, it is not clear at this time whether an NMR-based route or an X-ray based route may be the most efficient approach for functional assignment. Bottlenecks in both rapid assignment and rapid structure generation, as well as in high volume protein preparation still persist and these problems certainly need to be addressed before structural proteomics can become a reality.

2.6 References

1. Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J. R., Booth, V., Mackereth, C. D., Saridakis, V., Ekiel, I., Kozlov, G., Maxwell, K. L., Wu, N., McIntosh, L. P., Gehring, K., Kennedy, M. A., Davidson, A. R., Pai1, E. F., Gerstein, M., Edwards, A. M. and Arrowsmith, C. H. (2000) *Nature Struct. Biol.* 7, 903 – 909.

2. Skolnick, J., Feltrow, S. J. and Kolinski, A. (2000) *Nat. Biotech* 18, 283-287.

3. Brenner, S. E. and Levitt, M. (2000) *Protein Sci.* 9, 197-200.

4. Sali, A. (1998) *Nature Struct. Biol.* 5, 1029-1032.

5. Smith, D. R., Doucette-Stamm, L. A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., Harrison, D., Hoang, L., Keagle, P., Lumm, W., Pothier, B., Qiu, D., Spadafora, R., Vicaire, R., Wang, Y., Wierzbowski, J., Gibson, R., Jiwani, N., Caruso, A., Bush, D., and Reeve, J. N. (1997) *J. Bacteriol.* 179, 7135-7155.

6. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) *Nucleic Acids Res.* 25, 3389-3402.

7. Yee, A., Booth, V., Dharamsi, A., Engel, A., Edwards, A.M. and Arrowsmith, *C.H. (2000)) Proc. Natl. Acad. Sci. USA. 6, 6311-6315.*

8. Wishart, D.S., Bigam, C.G., Yao, J., Abilgaard, F., Dyson, H.J., Oldfield, E., Markley, J.L. and Sykes, B.D. (1995) *J. Biomol. NMR*, 6, 135-140.

9. Jeener, J., Meier, B.H., Bachmann, P., and Ernst, R.R. (1979) *J. Chem. Phys.* 71, 4546-4553.

10. Kay, L.E., Keifer, P., and Saarinen, T. (1992) J. Am. Chem. Soc., 114, 10663-10665.

11. Zhang, O., Kay, L.E., Olivier, J.P. and Forman-Kay, J.D., (1994) *J. Biomol. NMR,* 4, 845-858.

12. Kuboniwa, H., Grzesiek, S., Delaglio, F. & Bax, A. (1994) *J. Biomol. NMR*, 4, 871-878.

13. Pascal, S., Muhandiram, T., Yamazaki, T., Forman-Kay, J. D. & Kay, L. E. (1994) *J. Magn. Reson. 101, 197-201.*

14. Kay, L. E., Xu, G. Y. & Yamazaki, T. (1994) *J. Magn. Reson. Ser. A* 109, 129-133.

15. Grzesiek, S. & Bax, A. (1992) *J. Am. Chem. Soc.* 114, 6291-6293.

16. Kay, L. E., Xu, G. Y., Singer, A. U., Muhandiram, D. R. & Forman-Kay, J. *J. Magn. Reson. Ser. B* 101, 133-136.

17. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. & Bax, A. (1995) *J. Biomol. NMR. 66, 277-293.*

18. Fairbrother, W.J., Champe, M.A., Christinger, H.W., Keyt, B.A. and Starovasnik, M.A. (1997) *Protein Science*, 6, 2250-2260.

19. Clore, G. M., Bax, A., and Gronenborn, A. M. (1991) *J. Biomol. NMR.* 1, 13-22.

20. Wüthrich, K., Billeter, M., and Braun, W. (1983) *J. Mol. Biol.* 169, 949-961.

21. Gagne, S.M., Tsuda, S, Li, M.X., Chandra, M., Smillie, L.B. and Sykes, B.D. (1994) *Protein Science*, 3, 1961-1974.

22. Wishart, D. S, and Sykes, B. D. (1994) *Methods Enzymol.* 239, 363-392.

23. Brünger, A. T. (1993) *X-PLOR Manual*, version 3.1, Yale University, New Haven, CT.

24. Nilges, M., Clore, G.M. and Gronenborn, A.M. (1988) *FEBS Lett.* 229, 317-324.

25. Laskowski, R. A., Rullmann, J. A. C., MacArthur, M, W., Kaptein, R. and *Thornton, J. M. (1996). J. Biomol. NMR. 8, 477-486.*

26. Wishart, D.S., Willard, L. and Sykes, B. D. (1995) University of Alberta (http://redpoll.pharmacy.ualberta.ca)

27. Koradi, R., Billeter, M. and Wuthrich, K. (1980) *Biochem. Biophys. Res. Commun.* 95, 1-6.

28. Dyson, H. J. (1995) *Methods Enzymol.* 252, (293-306).

29. Dyson, H. J., feng, M. F., Tennant, L.L., Slay, I., Lindell, M., Cui, D. S., Kuprin, S. and Holmgren, A. (1997) *Biochemistry* 36, 2622-2236.

30. Pace, C. N., Vajdos, F., Fee, L., Grimsley, G., and Gray, T. (1995) *Protein Sci.* 4, 2411-2423.

31. Guex, N. and Peitsch, M.C. (1997) *Electrophoresis* 18, 2714-2723.

32. Hart, T. N., Ness, S. R. and Reid, R. J. (1997) *Proteins. Suppl* 1, 205-209.

33. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F. and Hermans, J. (1981) Interaction models for water in relation to protein hydration. *In* Intermolecular forces. *Edited by* B. Pullman. D. Reidel Publishing Company, Dordrecht, the Netherlands. pp. 331-342.

34. van Gunsteren, W. F., Daura, X. and Mark, A. E. (1998) *Encyclopaedia of Computational Chemistry* 2, 1211-1216.

35. Berendsen, H. J. C., van der Spoel, D. and van Drunen, R. (1995) Comp. Phys. Comm. 91, 43-56.

36. Vriend, G. (1990) *J. Mol. Graph.* 8, 52-56.

37. Laskowski, R. A., MacArthur, M. W., Moss, D. S. and Thornton, J. M. (1993) *Journal of Applied Crystallography*, 26, 283-291.

38. Lavigne, P., Bagu, J. R., Boyko, R., Willard, L., Holmes, C. F., Sykes, B. D., (2000) *Protein Sci.* 9, 252-264.

39. Holmgren, A. (1989*) J. Biol. Chem.* 264, 13963-13966.

40. Eklund, H., Gleason, F.K. and Holmgren, A. (1991) *Proteins* 11, 13-28.

41. Eklund, H., Cambillau, C., Sjoberg, B. M., Holmgren, A., Jornvall, H., Hoog, J. O. and Branden, C. I. (1984) *EMBO J.* 3, 1443-1449.

42. McFarlan, S. C., Terrell, C. A. and Hogenkamp, P. C. (1992) *J. Biol. Chem.* 267, 10561-10569.

43. Holm, L. and Sander, C. (1993) *J. Mol. Biol.* 233, 123-138.

44. Holm, L. and Sander, C. (1996) *Science*. 273, 595-602.

45. Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995). *J. Mol. Biol.* 247, 536-540.

46. Shindyalov, I. N. and Bourne, P. E. (1998) *Protein Eng.* 11, 739-747.

47. Lee, D. Y., Ahn, B-Y. and Kim, K-S. (2000) *Biochemistry*. 39, 6652-6659.

48. Takahashi, N. and Creighton, T. E. (1996) *Biochemistry*. 35, 8342-8353.

49. Gan, Z.R., Polokoff, M. A., Jacobs, J. W. and Sardana, M. K. (1990) *Biochem. Biophys. Res. Comm.* 168, 944-951.

50. Mieyal, J. J., Starke, D. W., Gravina, S. A. and Hocevar, B. A. (1991) *Biochemistry*. 30, 8883-8891.

51. Jeng, M. F. and Dyson, H. J. (1996) *Biochemistry*. 34, 611-619.

52. Dyson, H. J., Feng, M. F., Tennant, L. L., Slay, I., Lindell, M., Cui, D. S., Kuprin, S. and Holmgren, A. (1997) *Biochemistry*. 29, 4129-4136.

53. Sun, C. H., Berardi, M. J. and Bushweller, J. H. (1998) *J. Mol. Biol.* 280, 687-701.

54. Doublie, S., Tabor, S., Long, A. M., Richardson, C. C. and Ellenberger, T. (1998) *Nature*. 391, 251-258.

55. Ren B. Tibbelin G. de Pascale D. Rossi M. Bartolucci S. Ladenstein R. (1998) *Nature Struct. Biol.* 5, 602-611.

56. Nordstrand, K., Aslund, F., Holmgren, A., Otting, G. and Berndt, K.D. (1999) *J. Mol. Biol.* 286, 541-552.

57. Gan, Z. R., and Wells, W. W. (1987) *J. Biol. Chem.* 262, 6704-6707.

58. Vohnik, S., Hanson, C., Tuma, R., Fuchs, J. A., Woodward, C. and Thomas, G.J., Jr. (1998) *Protein Sci.* 7, 193-200.

59. Li, H., Hanson, C., Fuchs, J. A., Woodward, C., and Thomas, G. J., Jr. (1993) *Biochemistry* 32, 5800-5808.

60. Grauschopf, U., Winther, J. R., Korber, P., Zander, T., Dallinger, P., and Bardwell, J. C. (1995) *Cell* 83, 947-955 .

61. Wang, Y. *Ph.D Thesis* (1999) University of Alberta, Edmonton, AB, Canada.

62. Powles, R. E., Deane, S. M., and Rawlings, D. E. (1995) *Microbiology* 141, 2175-2181.

63. Lim, C. J., Sa, J. H., and Fuchs, J. A. (1996) *Biochim. Biophys. Acta* 1307, 13-16.

64. Pille, S., Assemat, K., Breton, A. M., and Clement-Metral, J. D. (1996) *Eur. J. Biochem.* 235, 713-720.

65. Holmgren, A. (1985) Annu. Rev. Biochem. 54, 237-271.

66. Yee, A., Booth, V.,Dharamsi, A., Engel, A., Edwards, A. M. and Arrowsmith, C. H. (2000) *Proc. Natl. Acad. Sci. USA* 97, 6311-6315.

67. Vogt, G., Woell, S. & Argos, P. (1997) *J. Mol. Biol.* 269, 631-643.

68. Thompson, M. J. and Eisenberg, D. (1999) *J. Mol. Biol.* 290, 595-604.

Table 2.1: Structural statistics of the 20 lowest energy structures of *MtH*895

Restraint type	**Number of restraints**
Total unambiguous NOE distances	862
Intra-residue	302
Sequential	171
Medium Range	152
Long range	237
Hydrogen bond constraints	46
Dihedral angles Φ_1	41
Backbone stereochemistry	**Percent residues in**
Most favorable regions	81.1%
Allowed regions	17.1%
Generously allowed regions	1.4%
Non-allowed regions	0.4%
Structure ensemble	**Pairwise rms deviations** (Å)
Backbone (all)	0.56
Heavy atoms (all)	1.42

Table 2.2: Protein structures similar to *MtH*895 derived from a CE database search.

Proteins	PDB code	Zscore	RMSD (Å)	Alignment length	Sequence length	Percent Identity
Thioredoxin complexed with TR* (E. coli)	1F6M	4.2	2.5	73	108	15.1
PDO** (P.furiosus)	1A8L	4.2	2.5	75	226	13
Thioredoxin (E. coli)	1XOB 1XOA 2TRX	4.2	2.5	73	108	16.4
Thioredoxin-M -Chloroplast (C. Reinhardtii)	1DBY	4.2	2.6	73	107	15.1
Thioredoxin-2 (Anabaena sp)	1THX	4.2	2.7	73	115	12.3
Thioredoxin (H. sapiens)	1AUC	4.2	2.8	73	105	12.3
Thioredoxin-M (Spinach Chlor-plast)	1FB6	4.2	2.8	73	105	12.3

*Thioredoxin reductase

**Protein disulfide oxidoreductase

Table 2.3: Comparative results of binding interactions of T7 DNA polymerase (pol) with *E.coli* thioredoxin (Trx), *MtH*895 and *E.coli* glutareodixn (Grx).

Structure	ΔH (kcal/mole)	$T\Delta S$ (kcal/mole)[e]	ΔG (kcal/mole)	K_a (M)	K_d (M)
T7pol+EcTrx(X-ray)[a]	3.74	27.53	-23.79	3×10^{17}	3×10^{-18}
T7pol+EcTrx[b]	3.99	28.25	-24.26	7×10^{17}	1×10^{-18}
T7pol+*MtH*895[c]	8.20	35.79	-27.62	2×10^{20}	5×10^{-21}
T7pol+EcGrx[d]	-1.48	3.81	-5.29	7×10^{3}	1×10^{-4}

a. X-ray crystal structure of T7 DNA pol bound to *E.coli* Trx (PDB: 1T7P)
b. Energy minimized structure of T7 DNA pol bound to *E.coli* Trx.
c. Energy minimized structure of T7 DNA pol bound to *MtH*895.
d. Energy minimized structure of T7 DNA pol bound to *E.coli* Grx.
e. T = 300 K .

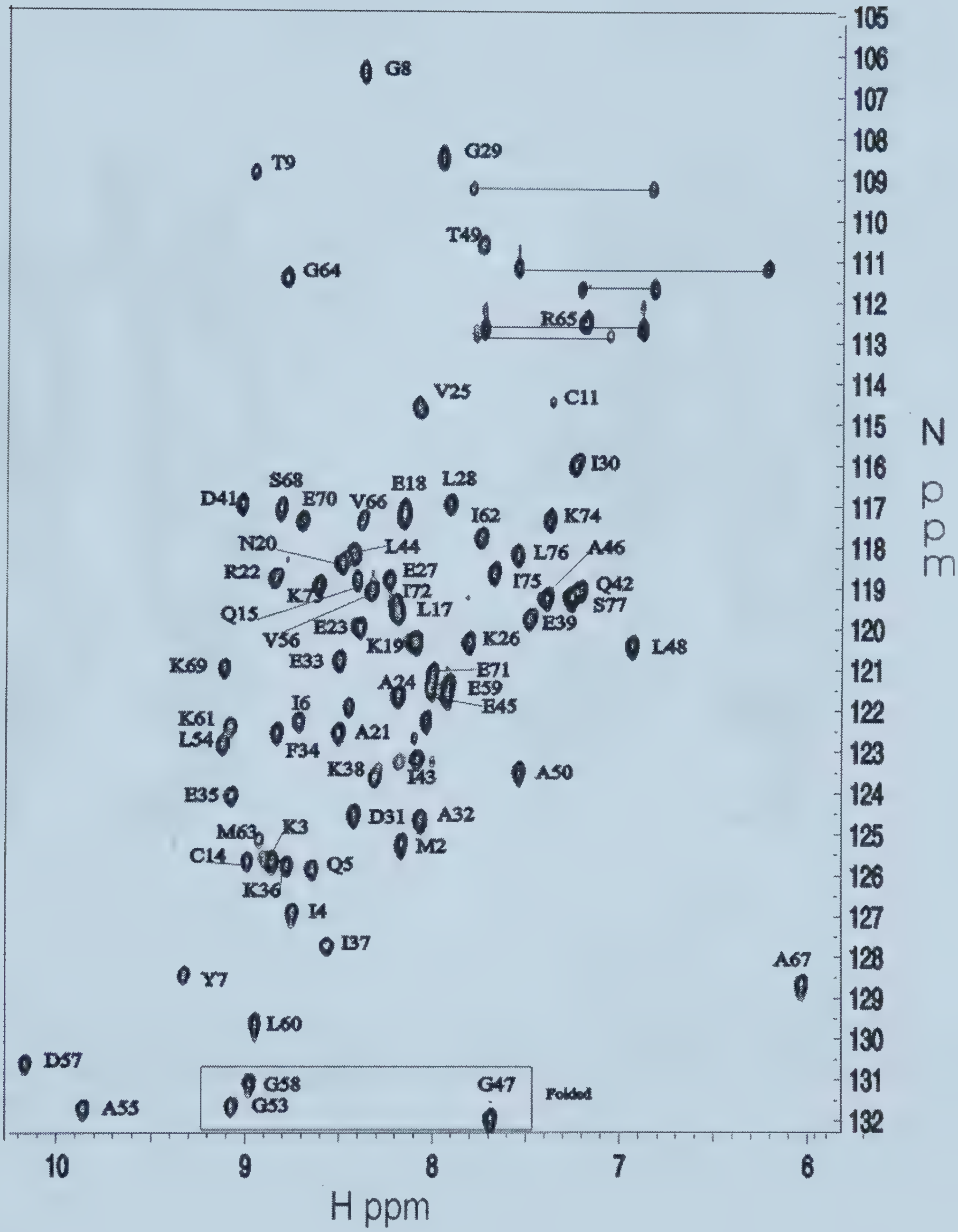


Fig. 2.1: ^{1}H-^{15}N HSQC spectrum from 1 mM *MtH*895 in 25 mM KH_2PO_4 , 200 mM NaCl , 0.5 mM EDTA , 20 mM DTT at pH7 collected at 25 ºC on a Varian Unity 500 MHz spectrometer.

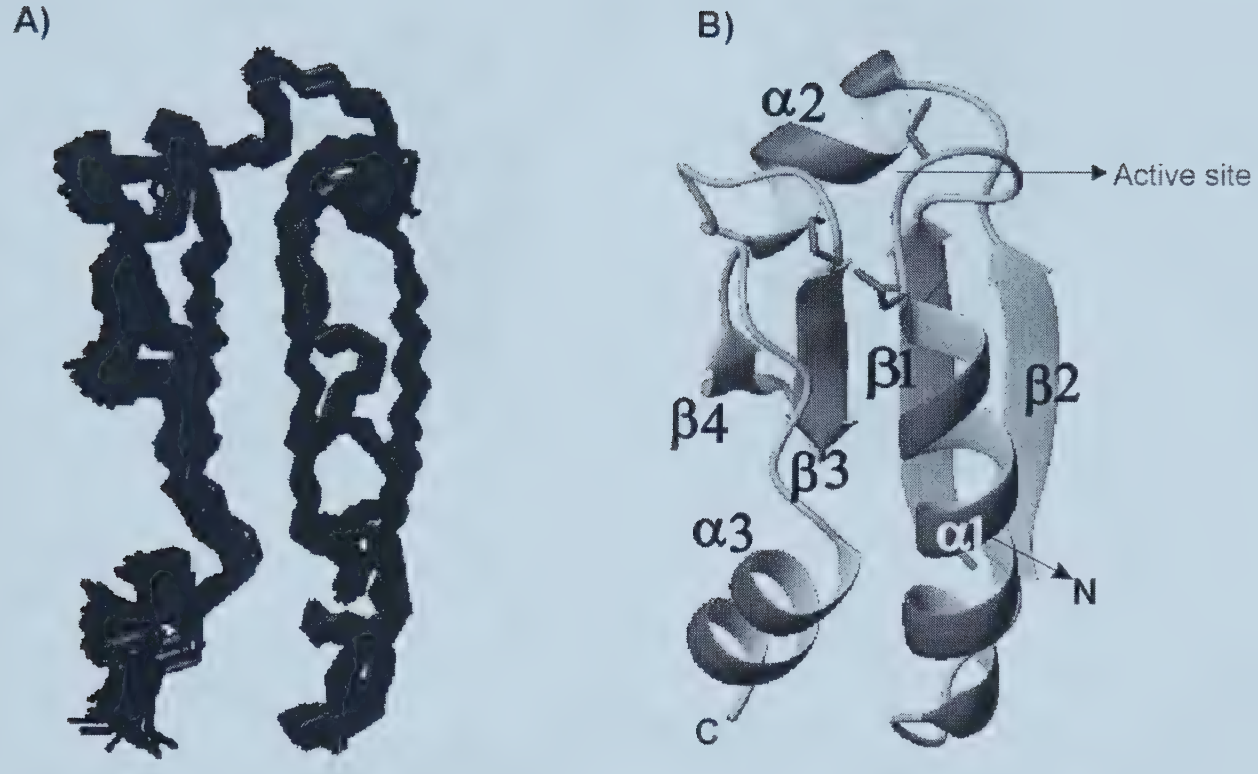


Fig. 2.2 (A): Ensemble of 20 minimum energy structures of *MtH*895, (B): Ribbon diagram of the representative structure of *MtH*895. The active site residues are shown with their side chains.

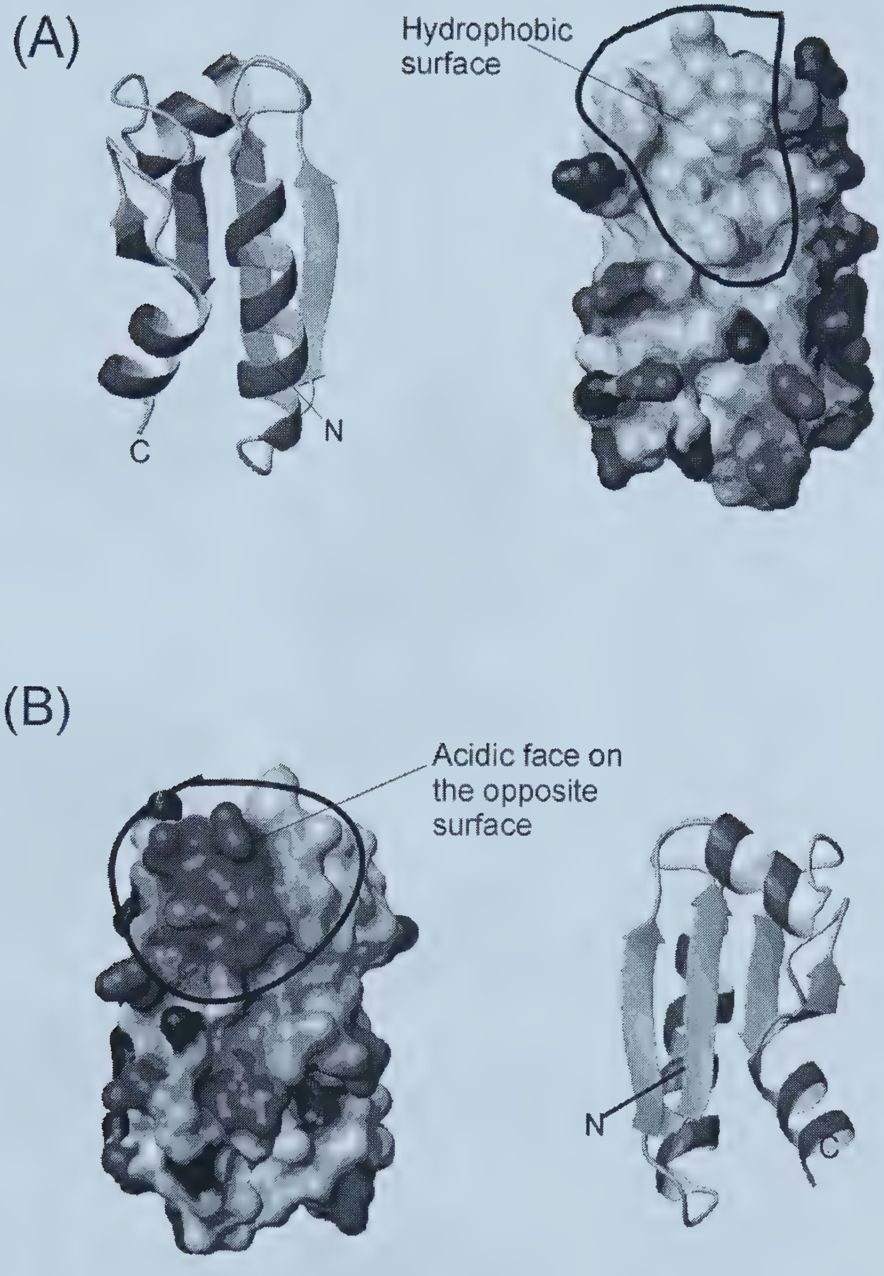
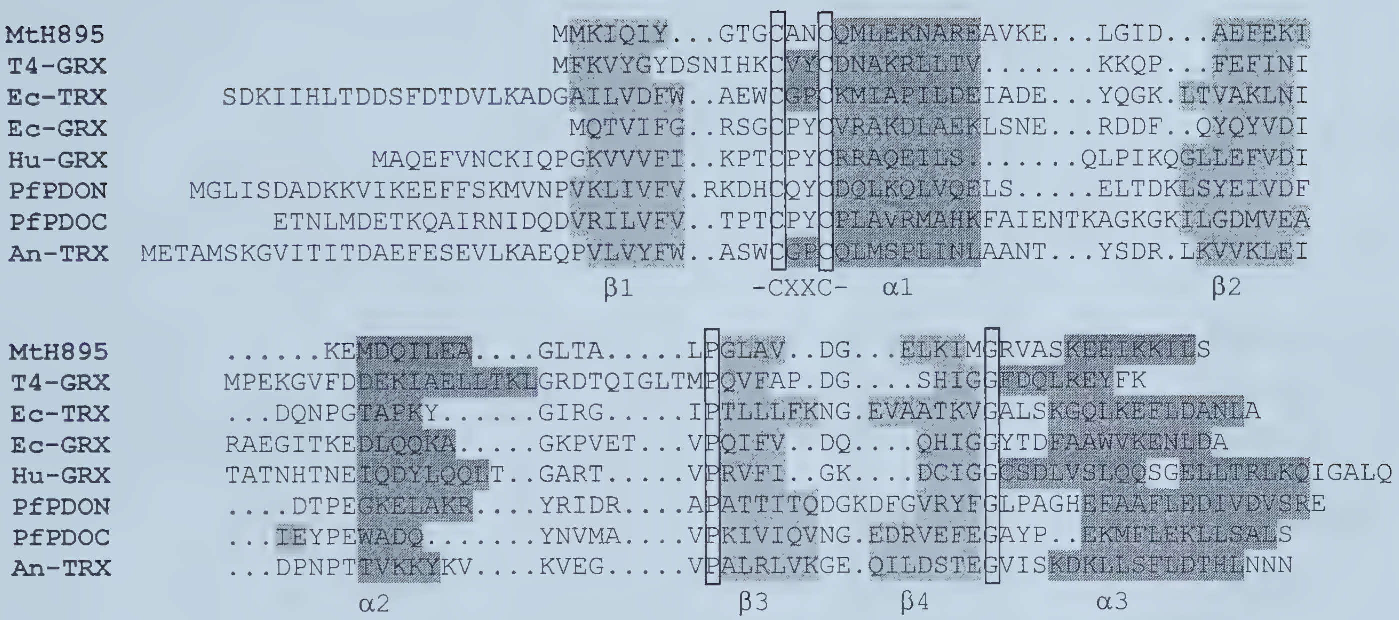


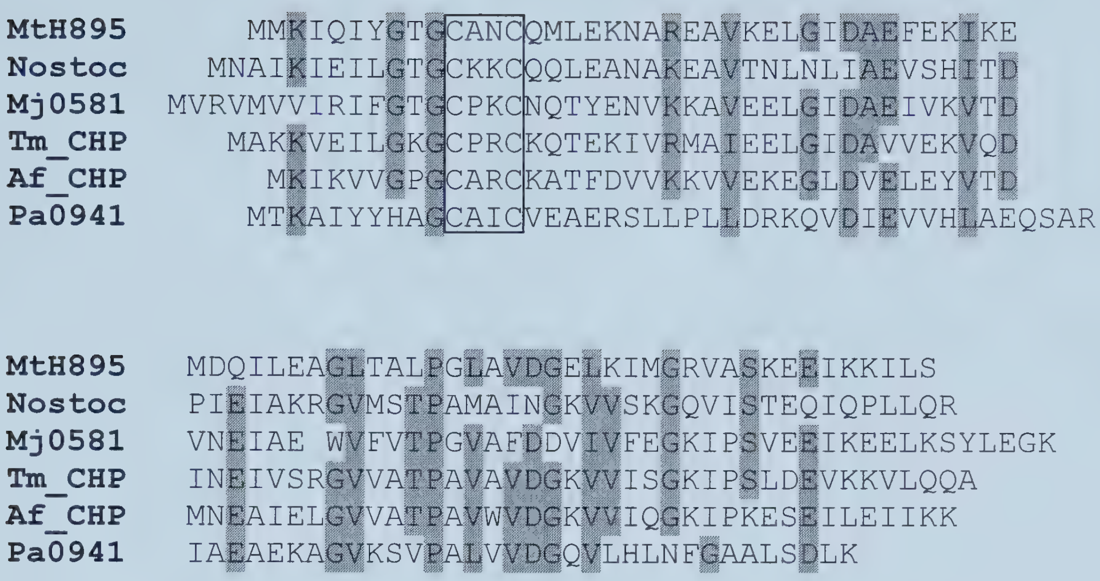
Fig. 2.3. Electrostatic potential surface of *MtH*895 showing the exposed hydrophobic surface on one side (A) and an acidic surface on the opposite face (B).



```
MtH895                             MMKIQIY...GTGCANCQMLEKNAREAVKE...LGID...AEFEKI
T4-GRX                             MFKVYGYDSNIHKCVYCDNAKRLLTV.......KKQP...FEFINI
Ec-TRX         SDKIIHLTDDSFDTDVLKADGAILVDFW..AEWCGPCKMIAPILDEIADE...YQGK.LTVAKLNI
Ec-GRX                              MQTVIFG..RSGCPYCVRAKDLAEKLSNE...RDDF..QYQYVDI
Hu-GRX                   MAQEFVNCKIQPGKVVVFI..KPTCPYCRRAQEILS.......QLPIKQGLLEFVDI
PfPDON       MGLISDADKKVIKEEFFSKMVNPVKLIVFV.RKDHCQYCDQLKQLVQELS.....ELTDKLSYEIVDF
PfPDOC            ETNLMDETKQAIRNIDQDVRILVFV..TPTCPYCPLAVRMAHKFAIENTKAGKGKILGDMVEA
An-TRX    METAMSKGVITITDAEFESEVLKAEQPVLVYFW..ASWCGPCQLMSPLINLAANT...YSDR.LKVVKLEI
                                      β1      -CXXC-  α1                  β2

MtH895    ......KEMDQILEA....GLTA.....LPGLAV..DG...ELKIMGRVASKEEIKKILS
T4-GRX    MPEKGVFDDEKIAELLTKLGRDTQIGLTMPQVFAP.DG....SHIGGFDQLREYFK
Ec-TRX    ...DQNPGTAPKY......GIRG.....IPTLLLFKNG.EVAATKVGALSKGQLKEFLDANLA
Ec-GRX    RAEGITKEDLQQKA.....GKPVET...VPQIFV..DQ....QHIGGYTDFAAWVKENLDA
Hu-GRX    TATNHTNEIQDYLQQLT..GART.....VPRVFI..GK....DCIGGCSDLVSLQQSGELLTRLKQIGALQ
PfPDON    ....DTPEGKELAKR....YRIDR....APATTITQDGKDFGVRYFGLPAGHEFAAFLEDIVDVSRE
PfPDOC    ...IEYPEWADQ.......YNVMA....VPKIVIQVNG.EDRVEFEGAYP..EKMFLEKLLSALS
An-TRX    ...DPNPTTVKKYKV....KVEG.....VPALRLVKGE.QILDSTEGVISKDKLLSFLDTHLNNN
                  α2                    β3       β4          α3
```

Fig. 2.4. Structure-based sequence alignment of *MtH*895, T4-glutaredoxin (T4-Grx), *E coli* thioredoxin (Ec-Trx), *E coli* glutaredoxin (Ec-Grx), Human glutaredoxin (Hu-Grx), *Pyrococcus furosius* protein disulfide oxidoreductase subunits N (PfPDON) and C (PfPDOC), *Anabaena* thioredoxin II (An-Trx).The α-helices and β-strands are marked in shades of grey. The 100% conserved residues are enclosed in boxes. The thioredoxins contain additional secondary elements in the N-terminal region which are not shown here.



```
MtH895       MMKIQIYGTGCANCQMLEKNAREAVKELGIDAEFEKIKE
Nostoc     MNAIKIEILGTGCKKCQQLEANAKEAVTNLNLIAEVSHITD
Mj0581   MVRVMVVIRIFGTGCPKCNQTYENVKKAVEELGIDAEIVKVTD
Tm_CHP      MAKKVEILGKGCPRCKQTEKIVRMAIEELGIDAVVEKVQD
Af_CHP        MKIKVVGPGCARCKATFDVVKKVVEKEGLDVELEYVTD
Pa0941       MTKAIYYHAGCAICVEAERSLLPLLDRKQVDIEVVHLAEQSAR


MtH895   MDQILEAGLTALPGLAVDGELKIMGRVASKEEIKKILS
Nostoc   PIEIAKRGVMSTPAMAINGKVVSKGQVISTEQIQPLLQR
Mj0581   VNEIAE WVFVTPGVAFDDVIVFEGKIPSVEEIKEELKSYLEGK
Tm_CHP   INEIVSRGVVATPAVAVDGKVVISGKIPSLDEVKKVLQQA
Af_CHP   MNEAIELGVVATPAVWVDGKVVIQGKIPKESEILEIIKK
Pa0941   IAEAEKAGVKSVPALVVDGQVLHLNFGAALSDLK
```

Fig. 2.5: Primary sequence alignment of the group of previously unidentified, highly conserved proteins from *M thermoautotrophicum* (*MtH*895), *Nostoc sp*, *M jannaschii* (*Mj*0581*), Thermotoga maritima* (Tm_CHP), *Archeoglobus fulgidus* (Af_CHP) and *Pseudomonous aeruginosa* (Pa0941). CHP stands for conserved hypothetical proteins. The active site –CXXC- motif is enclosed in a box. Highly conserved or similar residues are shaded.

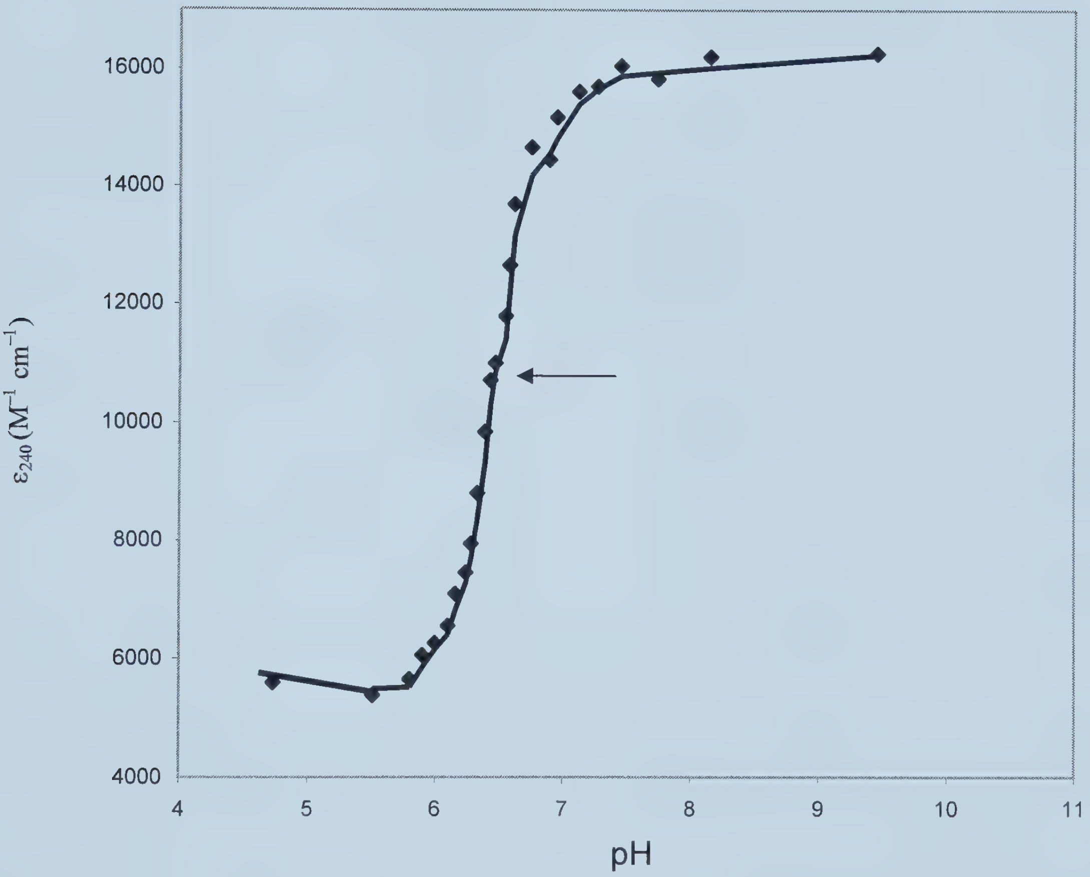


Fig.2.6: Measurement of thiol ionization in *MtH*895 by ultraviolet absorbance.

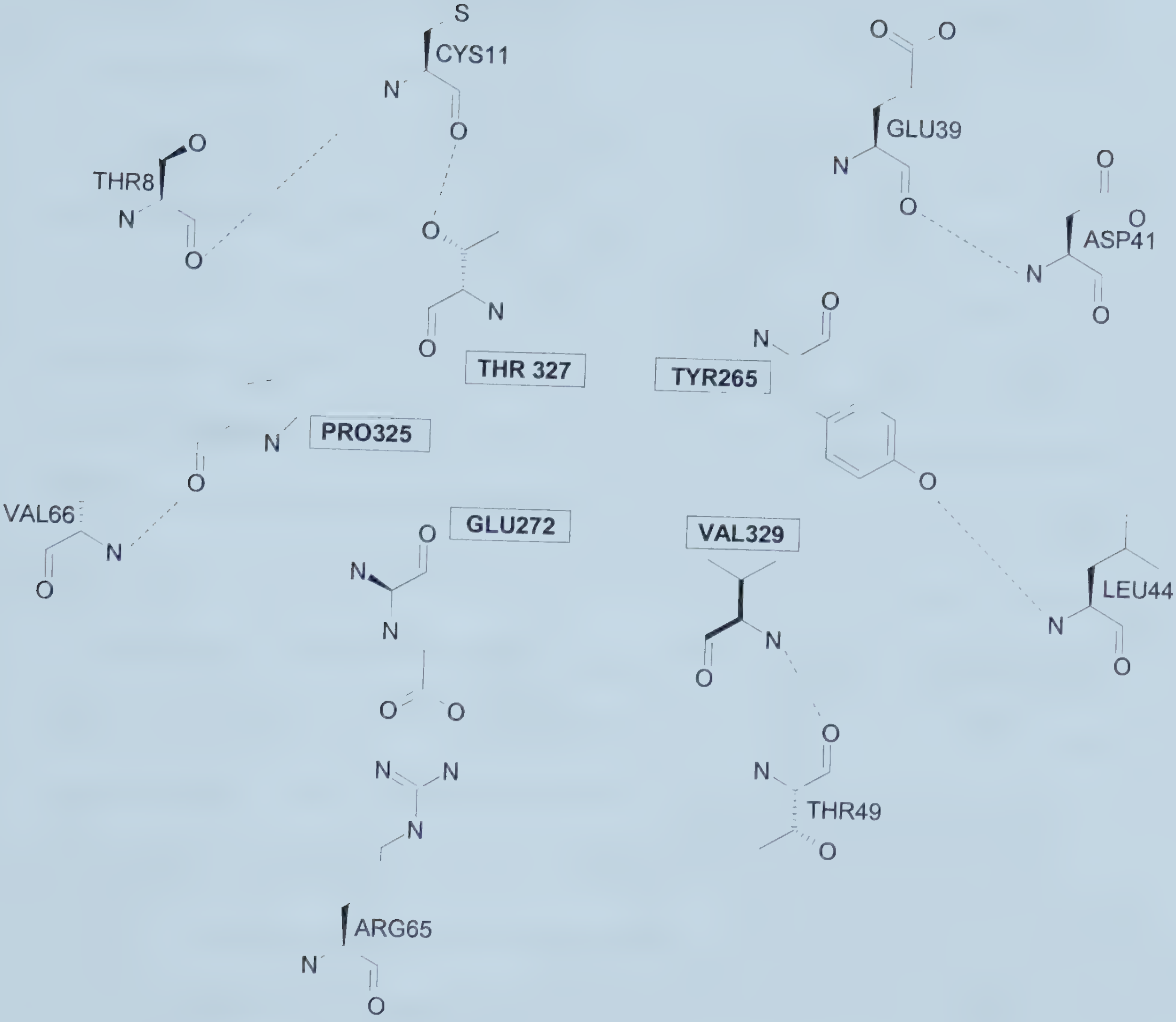


Fig. 2.7: Schematic representation of the interaction between T7 DNA polymerase and *MtH*895. Only a few interactions are shown here. Dashed lines represent H-bonds. Residues Cys11, Thr8, Val66, Arg65, Thr49, Leu44, Asp41, Glu39 belong to *MtH*895 while Thr327, Tyr265, Val329, Glu272, and Pro325 belong to T7 DNA polymerase.

Chapter 3

Expression, purification and isotopic labeling of *MtH*807---a putative thioredoxin from the archeon *M. thermoautotrophicum* strain *ΔH*

3.1 Introduction

Thioredoxin is a ubiquitous, multifunctional protein found in all living cells from archaebacteria to humans (*1,2*). Thioredoxins contain a well-conserved active site with two nearby cysteines (-Trp-Cys-Gly-Pro-Cys-). The oxidized form (thioredoxin-S_2) contains a disulfide bridge that is reduced to a dithiol by NADPH and the flavoprotein, thioredoxin reductase. The reduced form (thioredoxin-SH_2) is a powerful protein disulfide oxido-reductase. Depending on the chemical environment, thioredoxin can catalyze reduction of a protein disulfide bond or oxidation of a dithiol to a disulfide (*3*). Thioredoxins can also serve as reducing agents for ribinucleotide reductases, sulfates, and methionine sulfoxides (*1*). *E. coli* thioredoxin-SH_2 is an essential subunit of phage T7 DNA polymerase and is also required for the assembly of the filamentous viruses f1 and M13 (*4*). There is a large list of functional roles for thioredoxin in different biological systems involving regulation of protein activity by thiol redox control. For example, thioredoxins are important in the modulation of enzymes in the Calvin cycle (*1, 5*). They are involved in cell division (*8*) and other cellular processes such as protein assembly and repair, resistance to ionizing radiation, DNA replication and transcription. Thioredoxins have also been implicated to play a role in stimulating cancer cell growth as well as in apoptosis inhibition (*9*), thereby offering itself as a drug target towards treatment and prevention of cancer. Thioredoxins share many structural and functional similarities to another important

member of the thioltransferase superfamily—the glutaredoxins. However, while thioredoxins receive their electrons from NADPH via thioredoxin reductase, the transfer of electrons from NADPH to glutaredoxins proceed by the interaction of glutathione and glutathione reductase (*6*).

While much work has been done with thioredoxins from eukaryotes and eubacteria, little is known about the structure and function of thioredoxin/glutaredoxin-like proteins from archaebacteria. In this chapter we describe the expression, purification, isotopic labeling and preliminary NMR studies on a putative thioredoxin, *MtH*807 isolated from the archaeon *Methanobacterium thermoautotrophicum (ΔH).*

3.2 Materials and Methods

The plasmid DNA containing the *MtH*807 gene was a generous gift from the laboratory of Dr. Cheryl Arrowsmith of the University of Toronto. The *MtH*807 gene (Fig. 3.2, pg 88) was inserted into a T7 polymerase promoter driven vector, pET15b (Novagen), between the *Nde* I and *Bam*H I restriction sites. The construct contains a 2 kD linker and histidine-tag at the N-terminus of the recombinant protein (Fig. 3.1, pg 87).

Transformation

The pET15b plasmid was transferred into competent *Escherichia coli* cells [BL21(DE3) and BL21(DE3)pLysS strains] using standard electroporation methods. Specifically, *E. coli* cells were streaked onto an LB-agar plate and incubated

overnight at 37 °C. A culture flask containing 50 ml SOB was inoculated with a single bacterial colony from the LB-agar plate and placed in an incubator-shaker at 37 °C overnight. A 2 litre flask containing 500 ml SOB was inoculated with 5 ml of the overnight culture and incubated with vigorous agitation (250-300 rpm) at 37 °C for 2-3 hrs until the OD_{550} reached ~0.8. The culture was transferred into prechilled centrifuge bottles (250 ml) and centrifuged for 15 minutes at 3000-rpm (4 °C). The supernatant was discarded and the cells washed several times by resuspending them into ice-cold sterile distilled water followed by centrifugation for 15 minutes at 3000 rpm (4 °C). The cells were then washed twice in ice-cold sterile 10% glycerol and the supernatant carefully decanted. Finally, the cells were resuspended in a small volume of 10% sterile glyerol, dispensed in small aliquots (120 μl) in pre-chilled microcentrifuge tubes and stored at –80 °C.

The electroporation step was performed on a EC 100 Electroporator (E-C Apparatus Corporation, New York, USA) using sterile disposable 1 mm *E. coli* Pulser Cuvettes (Bio-Rad). The electrocomponent cells were removed from the –80 °C freezer and thawed on ice. 40 μl cell aliquots were used for each standard electroporation. 3-4 μl of the plasmid DNA (~1 μg/μl in 10 mM Tris, pH 7.5) was mixed gently with 40 μl of the electrocompetent cells and quickly transferred into the gap of a pre-chilled electroporation cuvette. The cuvette was then subject to a brief pulse of 1800 volts. The cuvette was removed and immediately 1 ml of SOC broth was added to the cells which were then transferred into 10 ml culture tubes. The cells were then incubated for 1 hour at 37 °C with shaking at 225 rpm. The resulting cell culture was diluted (1/10) and 50-100 μl aliquots were then plated out on LB-agar

plates containing 50 mg/ml of ampicillin to select for the positive transformants. The plates were incubated overnight at 37 °C. Positive transformants were selected and verified for *MtH*807 gene insertion by carrying out small-scale expression studies followed by SDS-PAGE analysis.

Expression

Large scale expression (1 L) in both rich (LB) and minimal (M9) media were performed using standard protocols, which were optimized to support the highest level of expression and recovery of the protein. BL21(DE3) cells harboring the recombinant plasmid were streaked onto LB-agar plates containing ampicillin (50 μg/ml). For BL21(DE3)pLysS cells the plates and the media also contained chloramphenicol (50 μg/ml). Chloramphenicol serves to maintain the pACYC-based plasmid carrying the T7 lysozyme gene. All plates were incubated overnight (12-16 hrs) at 37 °C. A single colony was then aseptically transferred to about 30 ml of media and grown overnight in an incubator shaker at 37 °C with vigorous shaking at 300 rpm. 10-15 ml of this starter culture was used to inoculate 1 L of media (starting OD_{600} = 0.1). When the OD_{600} reached ~0.5 (after 2-3 hrs of growth) IPTG was added to a final concentration of 1 mM in order to induce over-expression of the *MtH*807 gene. The cell culture was then allowed to grow at a reduced temperature (30 °C) for another four hours. The induction time was previously optimized by collecting culture samples at every hour after induction and analyzing the level of expression at each time point by SDS-PAGE. The cells were then harvested by

centrifugation at 5000 rpm for 15 minutes. The supernatant was discarded and the cell-pellet stored overnight at –20 °C.

The next day, the cell-pellet was thawed on ice and resuspended in 50 ml of lysis buffer (50 mM NaH_2PO_4, 300 mM NaCl, and 10 mM imidazole; pH 8.0). Lysozyme was added to a final concentration of ~80 μg/ml and the cell suspension incubated on ice for about 30 minutes to lyse the cell walls. The chromosomal DNA was broken up by sonication using three 20 seconds bursts (with 5 second pause intervals), which reduced the viscosity of the solution considerably. The lysate was then centrifuged for 1.5 hours at 15,000 rpm to remove all cell debris and the supernatant collected. About ten drops of 10% polyethylenimine was added to precipitate remaining polynucleic acids and the lysate was spun down for another 30 minutes at 12,000 rpm. The supernatant was collected again and incubated in a water bath at 70 °C with intermittent mixing for 12-15 minutes. This heating step precipitated most of the *E. coli* proteins. After another round of centrifugation at 10,000 rpm (20 minutes), the supernatant was loaded onto a Ni-NTA column for final purification.

For isotopic labeling and expression in minimal (M9) media an essentially identical procedure was followed. The composition of the minimal media used for optimum expression of *MtH*807 is presented in Table 3.1 (pg 84). The time of induction and growth period after induction were the same as that optimized for rich media. The overnight starter culture was also grown in minimal media (30 ml).

Purification

The recombinant *MtH*807 construct had a His_6-tag fused at its N-terminus in order to facilitate affinity purification. A Nickel-NTA affinity column was set up by packing a 10 ml syringe with ~4 ml of the Ni-NTA resin (Qiagen). The column was first equilibrated with 2 column volumes of lysis buffer (50 mM NaH_2PO_4, 300 mM NaCl, and 10 mM imidazole; pH 8.0) after which the protein supernatant (~30 ml) was loaded and allowed to flow through by gravity. The column eluant was monitored by UV absorbance at 280 nm using a single-path UV monitor. After sample loading was complete the column was subject to extensive washing with 250 ml of wash buffer (50 mM NaH_2PO_4, 300 mM NaCl, and 20 mM imidazole; pH 8.0). The washing step was performed at a flow rate of 0.2 ml/min, overnight, using a peristaltic pump. The next day the protein was eluted from the column using an elution buffer (50 mM NaH_2PO_4, 300 mM NaCl, and 250 mM imidazole; pH 8.0) at a flow rate of 0.4 ml/min.

The column eluant was collected in 1.5 ml fractions and the presence of protein in each fraction was monitored by UV absorbance (280 nm) using the first fraction as the reference. Each fraction was further analyzed by sodium dodecyl sulfate polyacrylamide gel electrophoresis (*12*). The fractions, which showed significant *MtH*807 bands, as judged with reference to a protein molecular weight standard, were pooled together and subject to a thrombin treatment to remove the His_6 tag. A schematic representation of the purification protocol is presented in Fig. 3.3 (.pg 89).

Removal of the His_6-tag

The His_6-tag (Fig. 3.1, pg 87) on the *MtH*807 protein construct was removed via proteolytic cleavage using thrombin. Thrombin recognizes the -Leu-Val-Pro-Arg-Gly-Ser- sequence found in the linker and cleaves on the C-terminal side of Arg. The pooled protein fractions were first dialyzed against phosphate buffer (20 mM KH_2PO_4 and 150 mM NaCl, pH 6.5) as the proteolytic activity of thrombin is four times more potent in phosphate buffer than in-stock elution buffer. The protein was then concentrated down to 900 μl using Centricon (Millipore) centrifugal filter units (molecular weight cut-off limit, 3000 daltons). 100 μl of 10X thrombin cleavage buffer (200 mM Tris-HCl pH 8.4, 1.5 M NaCl, 200mM NaCl, 25 mM $CaCl_2$) and 10 μl biotinylated thrombin, (Novagen) were added to the concentrated protein sample and incubated at room temperature for 16 hours. After the cleavage reaction, biotinylated thrombin was quantitatively removed with a Streptavidin Agarose slurry (supplied in the Novagen kit) using 16 μl settled resin per unit of thrombin. Biotinylated thrombin bound to the streptavidin agarose and the target protein was recovered by spin-filtration. SDS-PAGE was run to ensure that the His_6- tag was removed successfully.

^{15}N uniformly labeled MtH807

*MtH*807 was uniformly labelled with ^{15}N by growing the BL21(DE3)pLysS cells containing the recombinant plasmid in 1 L of M9 minimal media in which 1 g of unlabeled NH_4Cl was substituted with 1 g of 98% pure $^{15}NH_4Cl$ (Martek Biosciences

Corporation, Columbia, MD). *E. coli* cells utilize $^{15}NH_4Cl$ similarly to unlabeled NH_4Cl and therefore no changes in the growth or purification protocol were required.

$^{13}C/^{15}N$ uniformly labeled MtH807

In order to make $^{13}C/^{15}N$ isotopic labeling cost-effective, several rounds of expression using unlabelled M9 minimal media were conducted to determine the minimal amount of glucose (per liter batch) needed to successfully grow the cells. These studies (data not shown) indicated that at least 1.5 g of glucose per litre of minimal media was needed to ensure sufficient expression (~10 mg/L). Therefore, ^{13}C and ^{15}N isotopic labeling was performed by growing the BL21(DE3)pLysS cells in 1 L of minimal media with 1.0 g of $^{15}NH_4Cl$ and 1.5 g of $^{13}C_6$-glucose. As before, the BL21(DE3)pLysS cells utilized $^{13}C_6$-glucose similarly to the unlabelled glucose and so the expression and purification procedure remained unchanged.

*MtH*807 *quantification and storage*

The concentration of *MtH*807 was calculated by dividing the OD_{280} value by 0.29. The molar absorptivity was predicted using the following equation (10):

$$\varepsilon\ (280\ M^{-1}cm^{-1}) = (\#Trp)(5{,}500) + (\#Tyr)(1{,}490) + (\#Cys)(125). \qquad [3.1]$$

To stabilize the protein and to reduce the possibility of disulfide homodimerization, 0.5 mM EDTA and 1 mM DTT were added to the protein solution. 10 μls of 3% sodium azide was also added to protect the sample from bacterial growth. The sample was purged with argon gas to keep the thiols in a reduced state and sealed with parafilm and stored at 4 °C. *MtH*807 was found to be stable for about one week under

these conditions. It was also found that a high NaCl concentration (400 mM) and a minimum pH of 6.0 was needed to maintain the sample in solution for longer periods of time.

NMR-sample preparation

For 2D ^{1}H NMR experiments, protein samples (~1.0 mM) were prepared in a phosphate buffer (pH 6.5) with 10% D_2O, added as a lock solvent. For the 3D NMR experiments, which were collected over a much longer period of time, the salt concentration was raised to about 400 mM in order to maintain the protein in solution. A small quantity (0.1 mM) of 2,2-dimethyl-2-silapentane-5-sulfonic acid (DSS) was added to the samples as a chemical shift reference (11). All NMR spectra were collected on the reduced form of the protein. Prior to data collection, all samples were purged with argon for 5 minutes and sealed with parafilm and a rubber cap.

NMR spectroscopy

Homonuclear experiments: All the 1D and 2D ^{1}H NMR experiments were performed on a Varian VXR 500 MHz NMR spectrometer equipped with a 5-mm triple resonance probe. The basic pulse sequence of Bax and Davis (*14*) was used to perform standard TOCSY experiments. Acquisition times were generally set to 0.2 s, relaxation delays were 2.0 s and spin lock (MLEV-17) mixing times were generally set to 50 ms. NOESY spectra were collected using essentially identical acquisition parameters as the TOCSY spectra with mixing times ranging between 80-150 milliseconds. Both the TOCSY and the NOESY spectra were collected with 256 t_1

increments and spectral widths of 6000 Hz in both dimensions. All experiments were collected at 25 °C. The spectra were processed by zero-filling each data matrix to 4K x 4K complex data points. All FIDs were multiplied by ~90°-shifted sine-bell squared weighting function, in both dimensions, prior to Fourier transformation.

Heteronuclear experiments: Uniform ^{15}N labeled *MtH*807 was used to collect ^{1}H-^{15}N HSQC (*15,16*) ^{1}H-^{15}N TOCSY-HSQC (*16*) and ^{1}H-^{15}N NOESY-HSQC (*16*)experiments at 25°C on a Varian Unity 500 MHz spectrometer equipped with a 5 mm pulsed field gradient, triple resonance probe. The Experiments were set-up and performed by Dr. Stephane Gagne at NANUC, University of Alberta. A total of 138(t1) x 32(t2) x 416(t3) complex data points were collected for the TOCSY-HSQC (mixing time 32 ms) and NOESY-HSQC (mixing time 75 ms) with spectral widths of 5499, 1399, and 6300 Hz in each dimension. The carrier frequency for ^{1}H and ^{15}N were positioned at 4.75 and 118.50 ppm respectively. Chemical shifts were directly (^{1}H) and indirectly (^{13}C, ^{15}N) referenced to DSS at 0.00 ppm (*11*). Prior to Fourier transformation, each 3D data set was zero-filled to 1024 x 64 x 1024 complex data points and apodized using a shifted squared sine-bell window function to increase resolution.

3.3 Results and Discussion

*MtH*807 *Expression*

The pET15b vector with the *MtH*807 gene insert was retained as our expression vector primarily because it is known to have a high efficiency of transformation transcription and translation due to its bacterophage T7 transcription

and translation signals. The host strains chosen for expression purposes [BL21(DE3) and BL21(DE3)pLysS] are specific for the pET system as they contain a chromosomal copy of the T7 polymerase gene. These cells are lysogens of bacteriophage DE3, a lambda phage derivative that has the immunity region of phage 21 and carries a DNA fragment containing the *lacI* gene. The *lacI* gene is an IPTG-inducible *lacUV5* promoter. Addition of IPTG to a growing culture of these lysogens induces T7 RNA polymerase, which in turn transcribes the target DNA in the plasmid. These specific features of the pET-15b expression system greatly simplified the whole process of expression and purification of *MtH*807.

In our hands, the level of expression of *MtH*807 in rich medium was as high as 40 mg/L (Fig. 3.4, pg 90). In minimal media, however, the expression levels varied between 7-10 mg/L. The level of expression was however much lower in case of double isotopic labeling which will be discussed, in detail, later.

Several steps in the culture and lysis protocol were optimized to maximise the level of expression of the target protein. During the early phases of this project, we used Terrific Broth (TB) as the rich culture medium. However, TB contains a high amount (12 g/L) of tryptone, which is a milk product containing lactose. Lactose is a natural inducer of the *lacUV5* promoter and so the basal level of expression of the target gene was always higher than usual when the cells were cultured in TB. This problem was reduced when LB broth (tryptone content: 10 g/L) was used as the rich medium. The lower content of yeast-extract in LB (5 g/L as opposed to 24 g/L in TB) may also have some unknown effect on the basal level of expression. The level of protein expression in M9 minimal media was found to increase significantly by

supplementing the media with a vitamin mixture, ferrous sulfate (as the source of iron) and trace amount of zinc sulfate (see Table 3.1, pg 84). Expression levels of the recombinant protein at 25, 30 and 37 °C and for 2-14 hours of cell growth were compared using SDS PAGE (data not shown). Optimal expression of *Mt*H807 occurred at 30 °C, with cell harvest occurring 4 hours after IPTG induction. Typically the yield of *MtH*807 was slightly higher when the *E. coli* cells were lysed by sonication as opposed to conventional freeze-thaw methods. Sonication helped to reduce the lysate viscosity (and hence the handling problems) by shearing the cellular DNA.

Despite our best efforts, we found that the level of expression of *MtH*807 was not consistently uniform. At one point, the BL21(DE3) host cells became “leaky”, with the basal level of expression of the target gene being unusually high with almost no change in the expression level being observed after the addition of IPTG. The total protein yield from these cells was extremely poor. The reason for this high basal level of expression was likely due to the early expression of the T7 polymerase gene from the *lac*UV5 promoter. To get around this problem, we transformed the pET15b vector containing the *MtH*807 gene insert into a high-stringency expression host cell line (BL21(DE3)pLysS). These cells carry the pLysS plasmid which is a pET-compatible plasmid that carries the T7 lysozyme gene, a natural inhibitor of T7 RNA polymerase. The pLysS plasmid had no effect on the growth rate of the cells in culture and the total yield of the recombinant protein was found to be similar to that produced by the best-producing BL21(DE3) cells.

*MtH*807 *purification*

The purification of *MtH*807 procedure was made particularly simple because of the presence of a His_6 fusion tag, at the protein's N-terminus. The His_6-tag sequence can bind divalent Ni^{2+} cations immobilized on the nickel-nitrilotriacetic acid (Ni-NTA) metal-affinity agarose column. Proteins that don't have this tag simply pass through the column matrix, leaving only the target protein behind. We found that the target protein could be recovered to near homogeneity using a single imidazole elution step. The imidazole competes with the His_6 tagged protein for Ni^{2+} binding sites and consequently forces the bound protein off the resin. Several important factors had to be considered in designing the purification protocol. Stabilizing agents such as EDTA, DTT or 2-mercaptoethanol could not be present in the lysis, wash or elution buffers. In particular, thiol-containing components react with Ni^{2+} to form a brown precipitate while EDTA will chelate the Ni^{2+}, thereby stripping the column of its active affinity group. Column "clogging" was prevented by reducing the viscosity of the lysate as much as possible prior to loading. Using sonication to disrupt the cellular DNA and heating the cell lysate to 70 °C for 12-15 minutes (to selectively precipitate the host cell proteins), were two important steps that allowed the costly Ni-column to be reused several times.

The last step in the *MtH*807 purification process involved the thrombin cleavage of the His_6-tag from the parent protein. Thrombin is a site-specific endoprotease that cleaves after Arg in the consensus recognition sequence Leu-Val-Pro-Arg-Gly-Ser. This same consensus sequence connected the His_6 tag to the N-terminus of *MtH*807. The removal of the His_6-tag from *MtH*807 is not essential for

most structural or functional studies. However in our case its presence appeared to affect the interpretation of some NMR spectra. The amount of thrombin needed to cleave the target protein was determined using pilot digests with small amounts of target protein. The use of excess thrombin can occasionally result in unwanted proteolysis at secondary sites.

Isotopic Labeling

It has been reported (13) that swapping from rich growth media, to minimal media, usually results in sub optimal growth. So a process of minimal media adaptation was undertaken by iteratively growing *E. coli* cells on minimal media while maintaining the selection pressure for the target gene by including the appropriate antibiotics. In this adaptation, a 5% inoculum grown on rich medium was inoculated into minimal media and grown for 24 hours. A 5% inoculum from this culture was then passed to fresh minimal media and grown for another 24 hours. This passaging was repeated several times until a pilot 1 L expression study was carried out with unlabeled minimal medium to ensure that good expression levels of the target protein were maintained. 8% glycerol stocks of this pilot culture (collected prior to induction), were made and stored at –80 °C freezer for subsequent use in isotopic labeling studies.

Fig. 3.8 (pg 94) displays a ^{1}H-^{15}N HSQC spectrum collected on an ^{15}N labeled sample. As can be seen in this figure the spectra look very promising, with >96% of the expected peaks being visible. Preliminary spin system assignments of *MtH*807

based on the ^{1}H-^{15}N HSQC, TOCSY-HSQC and NOESY-HSQC spectra are presented in Table 3.2 (pg 85).

The yield of $^{15}N/^{13}C$ double labeled protein was much lower than expected (<3mg/L) (Fig. 3.6, pg 92). Fig. 3.7 (pg 93) shows a 1D ^{1}H spectrum collected on a ~0.01mM double labeled sample of *MtH*807 at pH 6.5. The unusually low expression of $^{15}N/^{13}C$ labeled protein was probably due to substantial loss of protein during thrombin treatment. However, the presence of low levels of $^{13}C_6$-glucose in the substituted media may have contributed to the low yield as well. A second round of pilot expression studies using varying amounts of glucose per liter of minimal media indicated that a minimum of 2.5 g/L of glucose is needed to express the protein in sufficient amounts for a NMR sample preparation. However, there may be other unknown factors which contributed to the low yield. No heteronuclear 3D experiments could be collected on the double labeled sample due to its low concentration as is evident from the 1D ^{1}H spectra.

3.4 Conclusions

High yield expression, purification and isotropic labeling of target proteins is a key factor in any detailed NMR study of proteins. In this project, *MtH*807 was successfully transformed, expressed, purified and ^{15}N labeled in sufficient quantities (~7-10 mg/L minimal media) to yield ‘good’ quality 2D and 3D NMR spectra. In the case of ^{15}N isotropic labeling, the minimum level of > 90% homogeneous labeling and >95% substitution in the target protein was satisfactorily met as evident from the 2D ^{1}H-^{15}N HSQC spectra. However, our attempt to produce a $^{15}N/^{13}C$ double-labeled

sample was not as successful due to relatively low expression levels and substantial losses from target peptide cleavage. Nevertheless, we were able to collect sufficiently good NMR spectra to begin the process of spectral assignment. Several pilot studies undertaken to optimize the expression and the purification were critical to the overall success of the project.

3.5 References

1. Eklund, H., Gleason, F.K. and Holmgren, A. (1991) *Proteins* 11, 13-28.

2. Holmgren, A. (1985) *Annu. Rev. Biochem*. 54, 237-271.

3. Prinz, W. A., Aslund, F., Holmgren, A. and Beckwith, J. (1997). *J Biol Chem.* 272, 15661-15667.

4. Lee, D. Y., Ahn, B-Y. and Kim, K-S. (2000) *Biochemistry*. 39, 6652-6659.

5. Thelander, L., and Reichard, P. (1979).) *Annu. Rev. Biochem*. 48, 133-158.

6. Holmgren, A. (1989) *J Biol Chem.* 264, 13963-13966.

7. McFarlan, S. C., Terrell, C. A. and Hogenkamp, P. C. (1992) *J. Biol. Chem.* 267, 10561-10569.

8. Pace, C. N., Vajdos, F., Fee, L., Grimsley, G., and Gray, T. (1995) *Protein Sci.* 4, 2411-2423

8. Buchanan, B.(1994). *Archives of Biochemistry & Biophysics*,:257-260.

9. Powis, G., Mustacich, D., Coon, A., Miranda-Vizuete, A., Damdimopoulos, A. E., Spyrou, G., Holmgren, A., Qin,J., Yang, Y., Velyvis, A. and Gronenborn, A. (2000) *Free Radic Biol Med* (29):312-322.

13. Mossakoska, D.E. and Smith, R.A.G. (1997) Methods in Molecular Biology Vol 60: Protein NMR Techniques, Reid, D.G. (ed), Humana Press Inc., Totowa, NJ, pp. 325-335.

14. Bax, A. and Davis, D.G. (1985) *J. Mag Res* 65: 355-360.

15. Kay, L.E., Keifer, P., and Saarinen, T. (1992) J. Am. Chem. Soc., 114, 10663-10665.

16. Zhang, O., Kay, L.E., Olivier, J.P. and Forman-Kay, J.D., (1994) *J. Biomol. NMR,* 4, 845-858.

Table 3.1: Reagents used for the expression of *MtH*807 in minimal media.

Reagents	**Quantity (per L)**
M9 salts	200 ml
Na_2PO_4	6.78 g
KH_2PO_4	3.00 g
NaCl	0.50 g
NH_4Cl	1.00 g
$MgSO_4$ (1.0 M)	2 ml
$FeCl_3$ (0.1 M)	10 µl
Vitamin Mixture	10 ml
Biotin	10mg/100 ml
Choline Chloride	“
Folic Acid	“
Niacinamide	“
D-pantothenic Acid	“
Pyridoxal	“
Riboflavin	1 mg/100 ml
Thiamine (0.1% w/v)	5 ml
$CaCl_2$ (0.01 M)	10 ml
$ZnSO_4$ (0.01 M)	5 ml
Ampicillin (50 mg/ml)	1 ml
Glucose (20% w/v)	20 ml

Table 3.2: Preliminary spin system assignments in *MtH*807

15N	NH	HA	HB	others	Probable amino acid
	7.28		2.88		
118.77	7.29	4.09	1.89	0.86, 0.77	Val
116.57	7.45	4.26	2.16	1.98	
118.12	7.50	4.36	2.70	2.49, 2.15, 1.20	
115.77	7.53	4.27			
118.41	7.60	5.31	4.01	1.23	Ser/Thr
120.29	7.62	5.13	4.32	1.67	Ser/Thr
125.73	7.64	4.43	0.80		Ala
107.35	7.67	4.08	2.23,		
117.34	7.72	4.09	1.09		Ala
122.39	7.74	5.03	1.19	3.22?	
120.15	7.80	4.01	2.58		
110.68	7.92		1.98		
121.71	7.95	3.89			
123.14	7.98	3.57		1.40	Ala
121.48	8.00	4.65			
118.90	8.03	4.17	1.20	0.85, 0.68	Val
116.75	8.04	3.79			
117.96	8.07	3.49			
120.81	8.11	4.18		2.18, 1.43	
126.35	8.17	4.14		0.92, 0.73	
121.81	8.22	5.39	1.96	1.59	
121.80	8.23	4.31			
111.54	8.24	4.14		3.40?	
122.44	8.27	4.95	2.58	2.83, 2.45	
112.35	8.29	4.37			
119.12	8.30	3.93, 3.92			Gly
121.43	8.33	4.48			
123.29	8.36	4.09	2.89,2.71		
121.70	8.38	4.61	1.50		Ala
118.27	8.39	3.89			
	8.45	4.38	2.07		
122.10	8.49	3.96		1.62?	
124.10	8.51	4.95		2.68, 2.37	
116.19	8.56	4.35			
120.80	8.60	3.74			
127.34	8.61	5.04	1.91	0.77, 0.70	Val
120.63	8.64	4.35	4.09	1.67	Ser/Thr
123.86	8.68	4.48	1.70	0.85	Val
128.78	8.76	4.39	4.09	0.65	Ser/Thr
	8.79		0.81	2.97, 2.60	
122.71	8.81	4.97			
128.55	8.82			0.86	
129.78	8.89	4.61	4.52		
	8.90	3.49			
121.35	9.02	4.35	1.84		

127.78	9.07	5.04			
128.22	9.18	4.14			
118.95	9.21	4.57	2.06	1.24	
	9.22		2.12		
112.20	9.25	5.00	2.24,	2.11	
119.84	9.34	5.08	2.11		
126.98	9.39	4.35, 4.05		0.81	
120.21	9.41	5.43	1.98		
128.19	9.95	4.65	1.29	0.90	
	9.97		1.94		
				2.67	

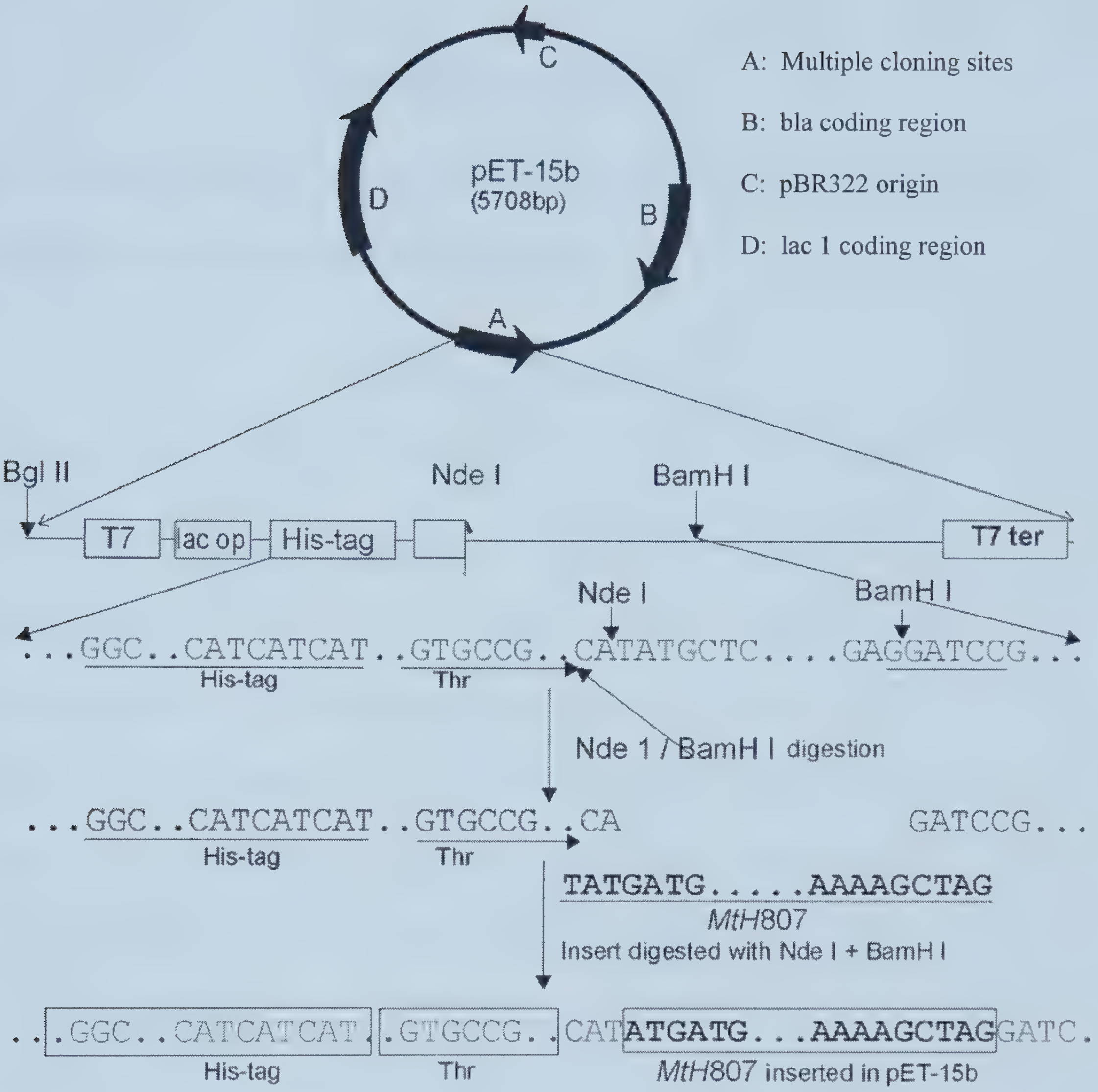

pET-15b
(5708bp)
A
B
C
D
A: Multiple cloning sites
B: bla coding region
C: pBR322 origin
D: lac 1 coding region
Bgl II
Nde I
BamH I
T7
lac op
His-tag
T7 ter
Nde I
BamH I
...GGC..CATCATCAT..GTGCCG..CATATGCTC....GAGGATCCG...
His-tag
Thr
Nde 1 / BamH I digestion
...GGC..CATCATCAT..GTGCCG..CA
GATCCG...
His-tag
Thr
TATGATG.....AAAAGCTAG
MtH807
Insert digested with Nde I + BamH I
...GGC..CATCATCAT..GTGCCG..CATATGATG...AAAAGCTAGGATC.
His-tag
Thr
MtH807 inserted in pET-15b

Fig. 3.1: Cloning of MtH807 construct into the vector pET-15b.

*MtH*807 protein sequence

MVVNIEVFTSPTCPYCPMAIEVVDEAKKEFGDKIDVEKIDIMVDREKAIEYGL

MAVPAIAINGVVRFVGAPSREELFEAINDEME

*MtH*807 gene sequence

```
1 atg gtt gtt aat ata gag gtt ttc aca tcc cca acc

  tgc cct tac tgt cca atg gca atc gag gtc gtt gat

  gag gcc aaa aag gaa ttc gga gac aaa att gat gtc

  gaa aag atc gac ata atg gtt gac cgg gaa aag gcc

  ata gag tat ggg ctg atg gct gtc cct gcc ata gca

  atc aac ggt gtt gtc agg ttc gtt ggg gcc cca agc

  agg gaa gaa ctc ttt gaa gcc ata aat gat gag atg

  gaa taa 258.
```

Fig. 3.2. *MtH*807 protein and nucleotide sequence. It is available from the GenBank database (http://www.ncbi.nlm.nih.gov) with the accession id GI:2621885.

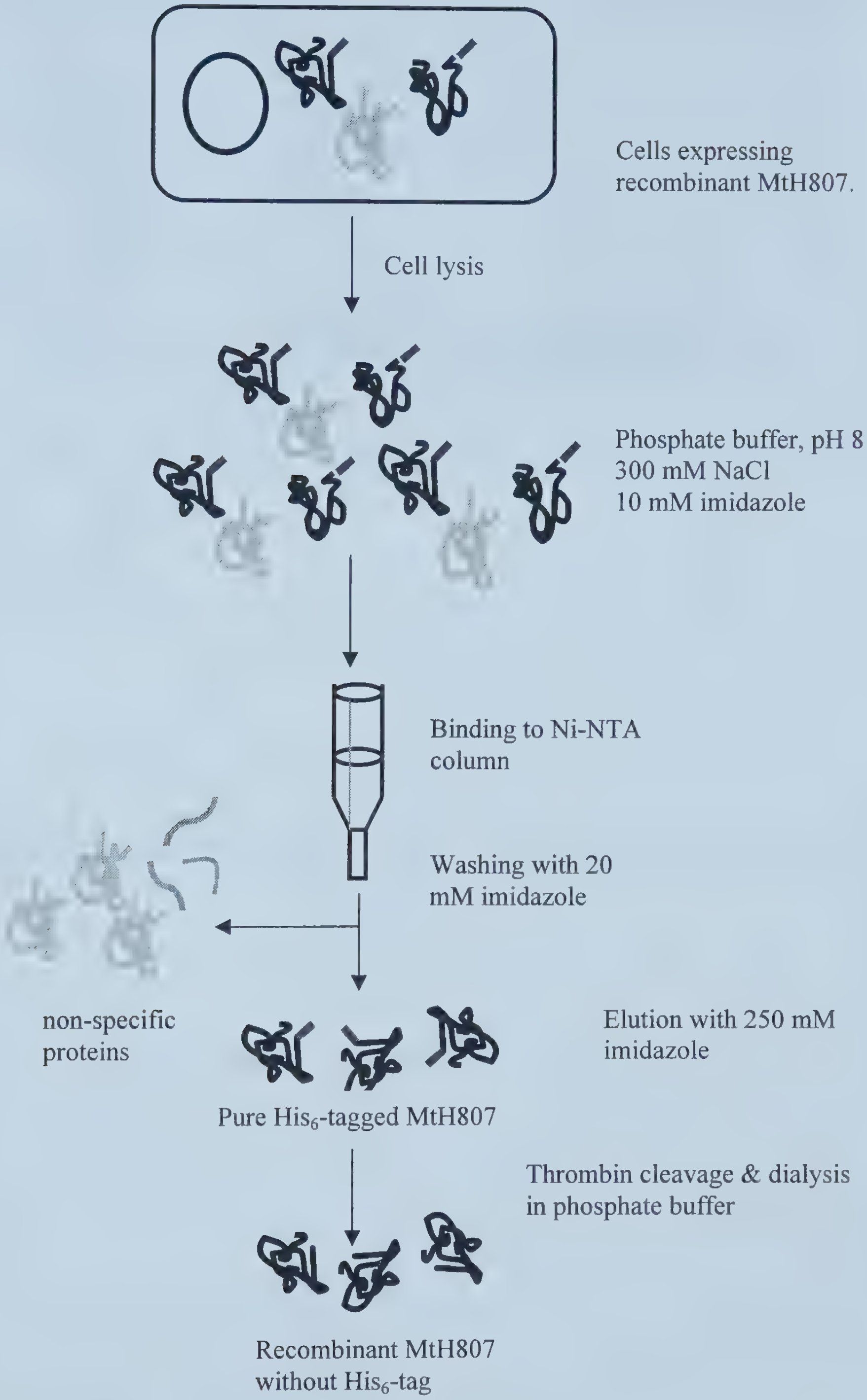

Cells expressing recombinant MtH807.
Cell lysis
Phosphate buffer, pH 8
300 mM NaCl
10 mM imidazole
Binding to Ni-NTA column
Washing with 20 mM imidazole
non-specific proteins
Elution with 250 mM imidazole
Pure His$_6$-tagged MtH807
Thrombin cleavage & dialysis in phosphate buffer
Recombinant MtH807 without His$_6$-tag

Fig. 3.3. Purification protocol of *MtH*807. BL21(DE3) cells expressing the recombinant *MtH*807 are first lysed in phosphate buffer, pH 8; supernatant loaded onto Ni-NTA agarose column where the protein binds; the protein is then eluted with imidazole which competes with His for binding to the column. Finally, the His-tag is removed by thrombin.

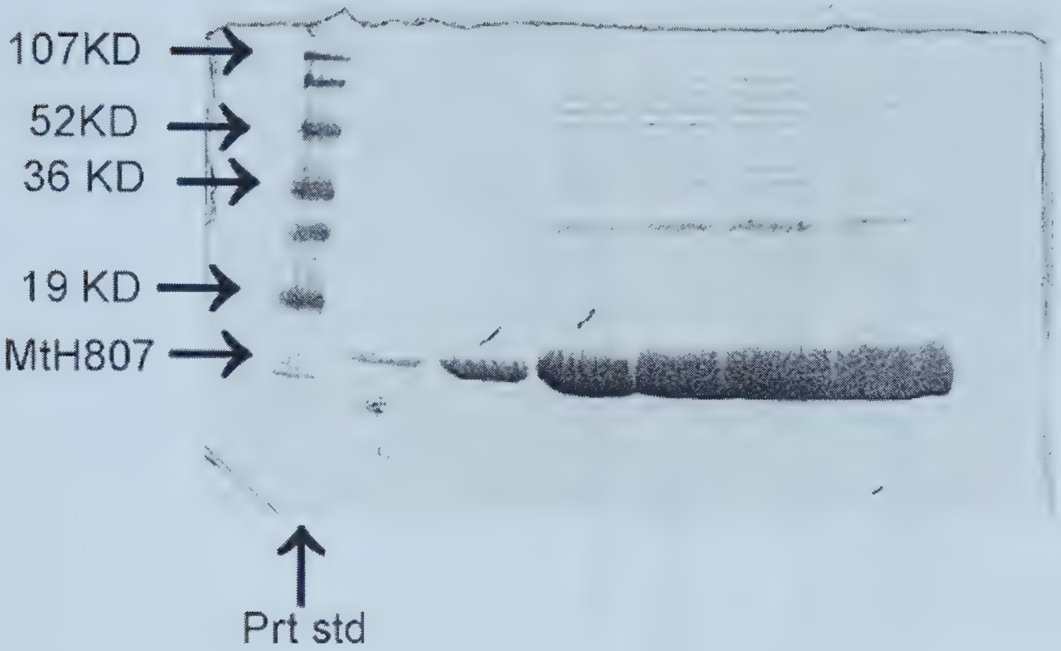


Fig.3.4: A polyacrylamide gel showing the level of expression of *MtH807* in rich medium.

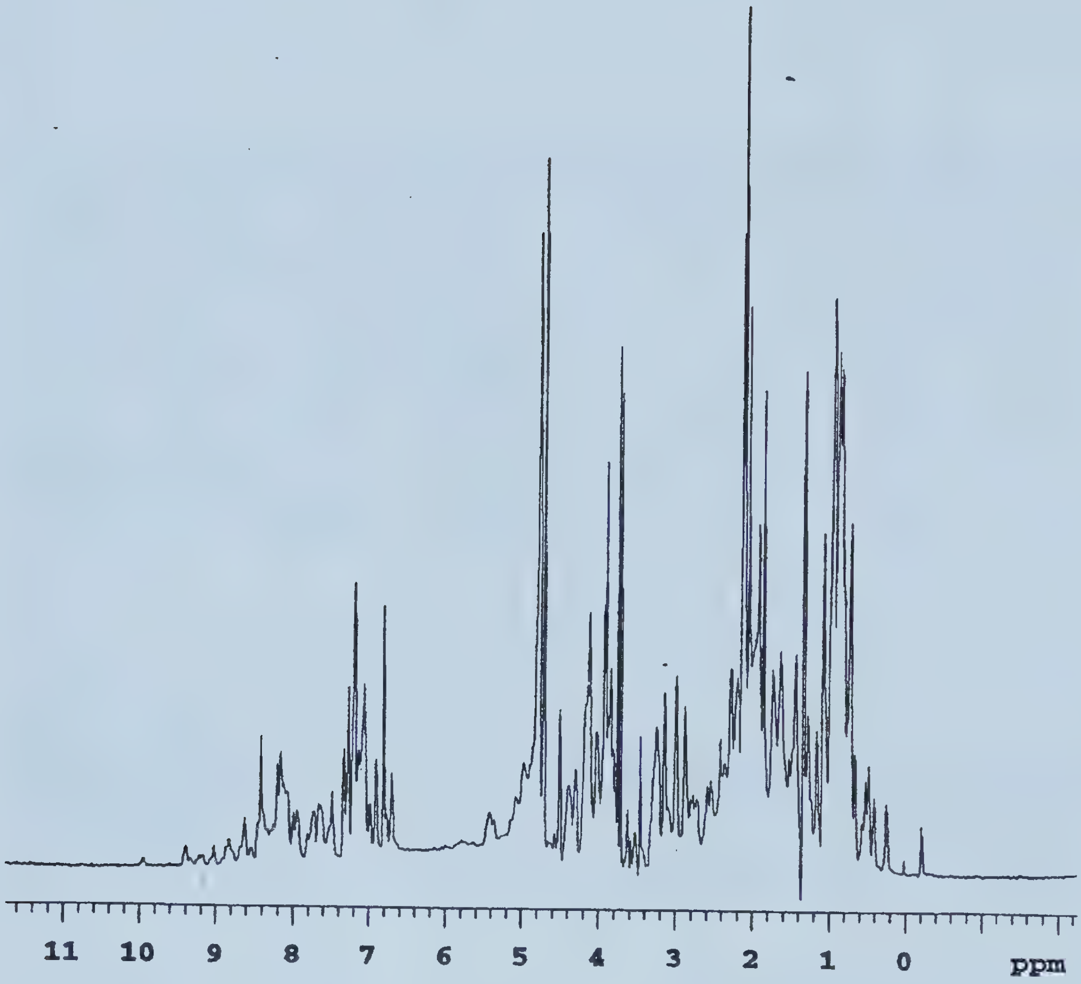


Fig.3.5: 1D proton NMR spectrum of *MtH807*.

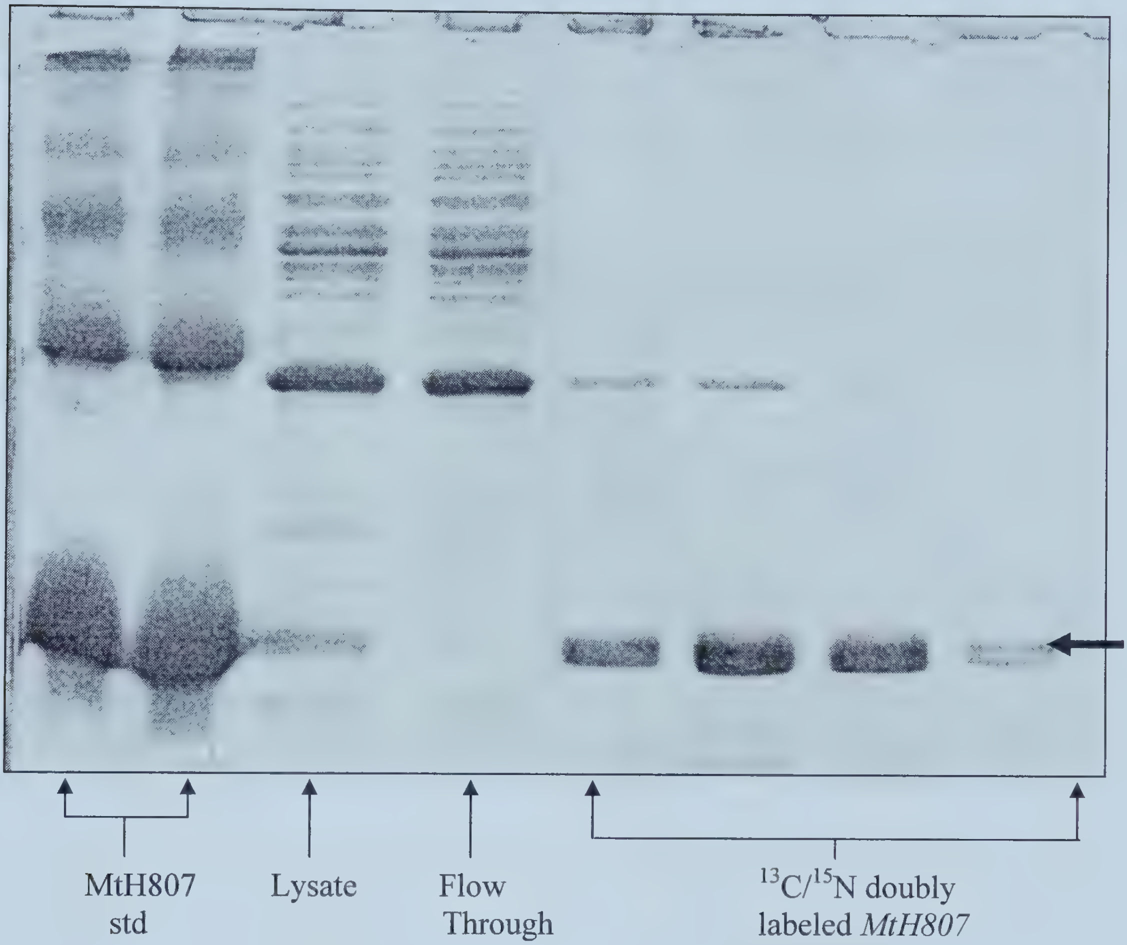
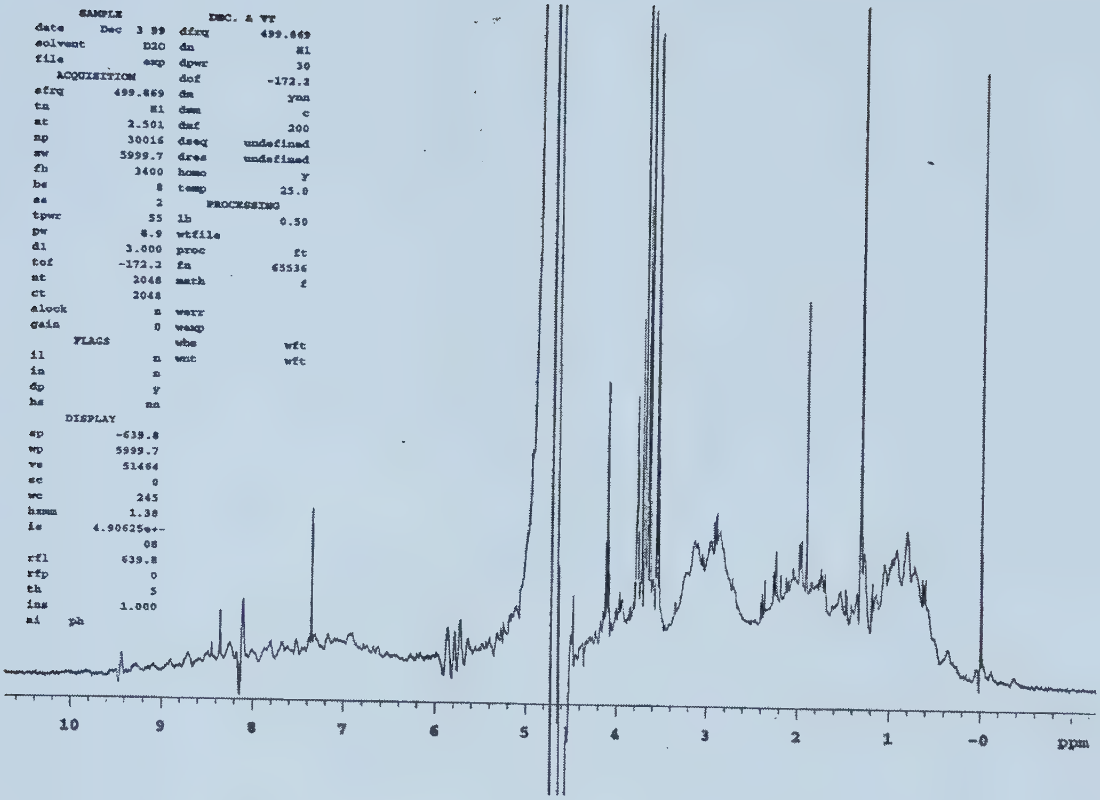


Fig. 3.6. SDS-PAGE showing expression of ^{13}C/^{15}N doubly labeled MtH807



```
        SAMPLE                     DEC. & VT
date       Dec  3 99   dfrq          499.869
solvent          D2O   dn                 H1
file             exp   dpwr               30
    ACQUISITION        dof            -172.2
sfrq         499.869   dm                ynn
tn                H1   dmm                 c
at             2.501   dmf               200
np             30016   dseq        undefined
sw            5999.7   dres        undefined
fb              3400   homo                y
bs                 8   temp             25.0
ss                 2         PROCESSING
tpwr              55   lb               0.50
pw               8.9   wtfile
d1             3.000   proc               ft
tof           -172.2   fn              65536
nt              2048   math                f
ct              2048
alock              n   werr
gain               0   wexp
       FLAGS           wbs               wft
il                 n   wnt               wft
in                 n
dp                 y
hs                nn
      DISPLAY
sp            -639.8
wp            5999.7
vs             51464
sc                 0
wc               245
hzmm            1.38
is         4.90625e+-
                  08
rfl            639.8
rfp                0
th                 5
ins            1.000
ai     ph
```

Fig. 3.7. 1D proton NMR spectrum of doubly $^{13}C/^{15}N$ doubly labeled *MtH807*.

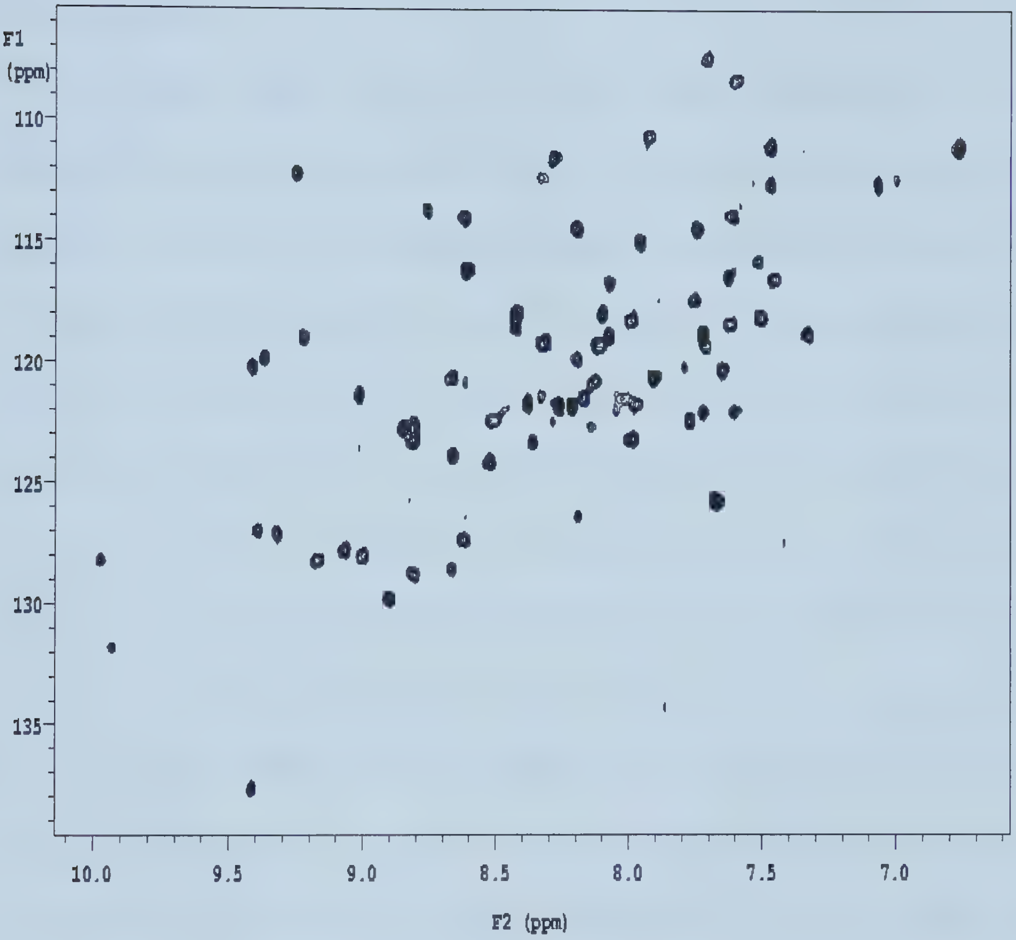


Fig.3.8: ^{15}N-HSQC spectrum of *MtH807* .

Chapter 4

General conclusions and future studies

4.1 Conclusions

This thesis centers on the application of NMR spectroscopy to structural proteomics. Structural proteomics is aimed at extracting functional clues to a large number of proteins by expressing, purifying and solving their structures in a high throughput manner. As part of a recently established structural proteomics initiative, our original intent was to go through the entire process of expressing, purifying, isotopically labeling and solving the 3D structure of a protein using NMR spectroscopy. We also intended to investigate the extent to which we could gain insight into an unknown protein's function through structural analysis. In chapter 2 of this thesis, we describe the determination of 3D structure of a previously unknown archeal protein (*MtH*895) using multidimensional heteronuclear NMR. Furthermore, through detailed structural analysis, functional assays and molecular docking studies we were able to show that even though *MtH*895 has a glutaredoxin-like structure it appears to function as a thioredoxin. In chapter 3 of this thesis, we describe the expression, purification, isotopic labeling, NMR sample preparation and acquisition of 2D and 3D heteronuclear NMR experiments on a putative thioredoxin, *MtH*807 isolated from the same archaeon, *M.thermoautotrophicum*. Thus, as described over these two chapters, we were able to successfully cover all the important aspects of a structural proteomics project.

As described in chapter 2, the quality of the NMR solution structure of *MtH*895 is comparable to a high-resolution (2.0 Å) X-ray structure. Before its

structure was solved it was classified as a protein of “unknown structure and function” with conserved sequence homologues in the archaeal and bacterial kingdom. After solving its structure, it initially appeared to resemble a glutaredoxin-like protein. However, structural and sequential comparisons to other known members of the thioredoxin/glutaredoxin family, measurements of its active site thiol pK_a values and molecular dynamics simulations of its interactions with T7 DNA polymerase suggested that, functionally, it is closer to a thioredoxin than a glutaredoxin. In fact, at just 77 residues it appears to be the smallest thioredoxin yet identified. We have also identified a group of previously unknown proteins from archaea and thermophilic bacteria that have high (34-44%) sequence identity with *MtH*895. These proteins have unique active site –CXXC- motifs not found in any known thioredoxin or glutaredoxin. We predict that these proteins form a new thioredoxin/glutaredoxin subclass bearing a close phylogenetic relationship to each other.

Chapter 3 describes the methods used to transform, express, purify, label ((^{15}N) and (^{15}N/^{13}C)) a second small thioredoxin-like protein from *M. thermoautotrophicum MtH*807. Several steps had to be optimized in order to maximize the yield of the protein. These included adding vitamin supplements to the M9 minimal media, changing host strains to reduce basal levels of expression and developing an appropriate growth adaptation protocol. The yield of ^{15}N labeled protein was more than enough to collect several ^{15}N-edited NMR experiments. However, the yield of doubly labeled (^{15}N/^{13}C) protein was not high enough to prepare a sufficiently concentrated (1 mM) NMR sample.

This may have been due to a combination of several factors including unanticipated losses during thrombin treatment and suboptimal ^{13}C-glucose content in media used for expressing the double labeled sample.

The combined success of the two projects suggests that structural proteomics is feasible and it further highlights the important role that NMR spectroscopy can play in it.

4.2 Future Studies

Now that we have determined *MtH*895 to be a thioredoxin, it should be possible to confirm our findings by carrying out the relevent biochemical experiments in the laboratory. For example, *MtH*895's thioredoxin activity could be confirmed by performing a DTNB-coupled reduction in the presence of *E. coli* thioredoxin reductase using *E. coli* thioredoxin as a positive control (1). Glutaredoxin activity could also be checked by carrying out the glutathione –disulfide transhydrogenase assay described by Gan et al. (2). If the result of the latter assay is negative whereas that of the former is positive it would conclusively prove our prediction that *MtH*895 is a thioredoxin. Interaction with T7 DNA polymerase could also be measured by a T7 bacteriophage plaque assay (3, 4). Unlike glutaredoxins, almost all known thioredoxins interact with T7 DNA polymerase. This interaction improves the processivity of T7 DNA pol complex about 1000-fold. Therefore T7 bacteriophage cannot survive in thioredoxin deficient bacteria. To test whether *MtH*895 can complement the thioredoxin deficiency, a thioredoxin deficient *E. coli* strain harboring a plasmid encoding *MtH*895 could be infected with T7 bacteriophage and

assayed for the number of plaques formed compared to the wild type *E. coli* strain. Thioredoxins reduce other proteins and is reduced, in turn, by an NADPH-dependent thioredoxin reductase in *E. coli*. It would be interesting to see if the putative thioredoxin reductase in *M.thermoautotrophicum* (*MtH*708) could reduce *MtH*895 with NADPH as a hydrogen donor or any other factor is involved in the thioredoxin system in the archaeon.

In case of the putative thioredoxin *MtH*807, the optimal protocol to express and purify the protein as well as the optimum solution conditions to collect NMR experiments have been established. The ^{15}N-NOESY and ^{15}N-TOCSY HSQCs collected on the protein along with the ^{15}N-HSQC spectrum could be analyzed further. This would go a long way towards completing the spin system identifications, backbone assignment as well as collecting NOE distance restraints. However, in order to expedite the process of assignments and structure generation, preparation of a ^{13}C/^{15}N double labeled sample of concentration at least 1 mM is recommended so that triple resonance experiments (HNCACB, CBCACONH) could be collected to complete the backbone assignments. Other experiments, such as the HCCH-TOCSY for side chain assignments, ^{15}N/^{13}C edited simultaneous NOESY-HSQCs for NOE, as well as experiments like HNHA or HMQC-J experiments for coupling constant determination could eventually be collected and analyzed for structure generation purposes. Similar biochemical assays (as mentioned in case of *MtH*895) could also be performed for *MtH*807 in order to confirm its putative function. It would be interesting to analyse the structural properties of both thioredoxins and be able to relate these structural differences in their functional

specificity. Moreover, these two proteins can also serve as "test" set to refine and optimize the novel, non-NOE structure generation approaches that are currently being developed in the lab that are being aimed towards high throughput protein structure generation. Last but not the least, the technical know-how acquired from these two projects could be applied to the study of more pharmaceutically relevent, disease-related proteins from more medically interesting organisms.

4.3 References

1. Holmgren, A. (1979) *J. Biol. Chem.* 254, 9113-9119.
2. Gan, Z.R. and Wells, W.W. (1986) *J. Biol. Chem.* 261, 996-1001.
3. Grauschopf, U., Winther, J.R., Korber, P., Zander, T., Dallinger, P. and Bardwell, J.C. (1995) *Cell* 83, 947-955.
4. Lee, D.Y., Ahn, B-Y and Kim, K-S (2000) *Biochemistry* 39, 6652-6659.

Appendix A

Chemical shift assignments of *MtH*895 in BMRB format

Atom #	Residue #	Residue Label	Atom Name	Atom Type	Shift/ ppm	Error/ ppm	Ambiguity Code
1	1	MET	H	H	8.18	0.05	1
2	1	MET	HA	H	4.43	0.05	1
3	1	MET	HB2	H	1.87	0.05	2
4	1	MET	HB3	H	1.87	0.05	2
5	1	MET	HG2	H	2.34	0.05	2
6	1	MET	HG3	H	2.26	0.05	2
7	1	MET	CA	C	55.10	0.05	1
8	1	MET	CB	C	33.29	0.05	1
9	1	MET	N	N	123.41	0.05	1
10	2	MET	H	H	8.17	0.05	1
11	2	MET	HA	H	4.56	0.05	1
12	2	MET	HB2	H	2.00	0.05	2
13	2	MET	HB3	H	1.85	0.05	2
14	2	MET	HG2	H	2.41	0.05	2
15	2	MET	HG3	H	2.16	0.05	2
16	2	MET	CA	C	56.06	0.05	1
17	2	MET	CB	C	35.10	0.05	1
18	2	MET	N	N	125.47	0.05	1
19	3	LYS	H	H	8.85	0.05	1
20	3	LYS	HA	H	4.85	0.05	1
21	3	LYS	HB2	H	1.69	0.05	2
22	3	LYS	HB3	H	1.58	0.05	2
23	3	LYS	HG2	H	1.29	0.05	2
24	3	LYS	HG3	H	1.17	0.05	2
25	3	LYS	HE2	H	2.80	0.05	1
26	3	LYS	CA	C	54.52	0.05	1
27	3	LYS	CB	C	33.95	0.05	1
28	3	LYS	N	N	125.73	0.05	1
29	4	ILE	H	H	8.73	0.05	1
30	4	ILE	HA	H	4.66	0.05	1
31	4	ILE	HB	H	0.92	0.05	1
32	4	ILE	HG12	H	0.70	0.05	2
33	4	ILE	HD1	H	0.43	0.05	1
34	4	ILE	CA	C	59.73	0.05	1
35	4	ILE	CB	C	38.47	0.05	1
36	4	ILE	N	N	126.88	0.05	1
37	5	GLN	H	H	8.62	0.05	1
38	5	GLN	HA	H	5.05	0.05	1
39	5	GLN	HB2	H	1.93	0.05	2
40	5	GLN	HB3	H	1.14	0.05	2
41	5	GLN	HG2	H	2.27	0.05	2
42	5	GLN	HG3	H	2.07	0.05	2
43	5	GLN	HE21	H	7.55	0.05	2
44	5	GLN	HE22	H	6.22	0.05	2
45	5	GLN	CA	C	54.92	0.05	1
46	5	GLN	CB	C	31.07	0.05	1

47	5	GLN	N	N	125.91	0.05	1
48	5	GLN	NE2	N	111.17	0.05	1
49	6	ILE	H	H	8.71	0.05	1
50	6	ILE	HA	H	4.94	0.05	1
51	6	ILE	HB	H	1.90	0.05	1
52	6	ILE	HG12	H	1.46	0.05	2
53	6	ILE	HG13	H	1.45	0.05	2
54	6	ILE	HG2	H	0.93	0.05	1
55	6	ILE	CA	C	58.49	0.05	1
56	6	ILE	CB	C	38.62	0.05	1
57	6	ILE	N	N	122.25	0.05	1
58	7	TYR	H	H	9.32	0.05	1
59	7	TYR	HA	H	5.27	0.05	1
60	7	TYR	HB2	H	2.62	0.05	1
61	7	TYR	HB3	H	2.50	0.05	1
62	7	TYR	HD1	H	6.72	0.05	1
63	7	TYR	HD2	H	6.72	0.05	1
64	7	TYR	HE1	H	6.55	0.05	1
65	7	TYR	HE2	H	6.55	0.05	1
66	7	TYR	CA	C	56.49	0.05	1
67	7	TYR	CB	C	42.03	0.05	1
68	7	TYR	N	N	128.52	0.05	1
69	8	GLY	H	H	8.37	0.05	1
70	8	GLY	HA2	H	5.00	0.05	2
71	8	GLY	HA3	H	3.90	0.05	2
72	8	GLY	CA	C	45.71	0.05	1
73	8	GLY	N	N	106.40	0.05	1
74	9	THR	H	H	8.99	0.05	1
75	9	THR	HA	H	4.64	0.05	1
76	9	THR	HB	H	4.46	0.05	1
77	9	THR	HG1	H	1.18	0.05	1
78	9	THR	HG2	H	1.18	0.05	1
79	9	THR	CA	C	62.12	0.05	1
80	9	THR	CB	C	69.15	0.05	1
81	9	THR	N	N	108.87	0.05	1
82	10	GLY	HA2	H	4.28	0.05	2
83	10	GLY	HA3	H	3.87	0.05	2
84	10	GLY	CA	C	45.74	0.05	1
85	11	CYS	H	H	7.35	0.05	1
131	11	CYS	HA	H	4.67	0.05	1
132	11	CYS	HB2	H	3.41	0.05	2
133	11	CYS	HB3	H	2.93	0.05	2
138	11	CYS	N	N	114.50	0.05	1
140	12	ALA	HA	H	4.55	0.05	1
141	12	ALA	HB	H	0.82	0.05	2
143	12	ALA	CA	C	51.72	0.05	1
144	12	ALA	CB	C	22.20	0.05	1
146	13	ASN	H	H	8.77	0.05	1
147	13	ASN	HA	H	4.53	0.05	1
148	13	ASN	HB2	H	3.11	0.05	1
149	13	ASN	HB3	H	2.23	0.05	1
150	13	ASN	HD21	H	7.78	0.05	2
151	13	ASN	HD22	H	7.06	0.05	2
153	13	ASN	CA	C	55.90	0.05	1
154	13	ASN	CB	C	37.49	0.05	1
156	13	ASN	N	N	118.50	0.05	1
157	13	ASN	ND2	N	112.84	0.05	1

158	14	CYS	H	H	9.01	0.05	1
159	14	CYS	HA	H	3.82	0.05	1
160	14	CYS	HB2	H	3.37	0.05	1
161	14	CYS	HB3	H	3.05	0.05	1
164	14	CYS	CA	C	63.99	0.05	1
165	14	CYS	CB	C	27.07	0.05	1
166	14	CYS	N	N	125.95	0.05	1
167	15	GLN	H	H	8.40	0.05	1
168	15	GLN	HA	H	4.06	0.05	1
169	15	GLN	HB2	H	2.06	0.05	2
170	15	GLN	HB3	H	2.00	0.05	2
171	15	GLN	HG2	H	2.34	0.05	2
172	15	GLN	HG3	H	2.33	0.05	2
173	15	GLN	HE21	H	7.21	0.05	2
174	15	GLN	HE22	H	6.82	0.05	2
176	15	GLN	CA	C	58.97	0.05	1
177	15	GLN	CB	C	28.57	0.05	1
180	15	GLN	N	N	118.87	0.05	1
181	15	GLN	NE2	N	111.70	0.05	1
182	16	MET	H	H	8.07	0.05	1
183	16	MET	HA	H	4.24	0.05	1
184	16	MET	HB2	H	2.08	0.05	2
185	16	MET	HB3	H	1.94	0.05	2
186	16	MET	HG2	H	2.63	0.05	2
187	16	MET	HG3	H	2.22	0.05	2
190	16	MET	CA	C	58.26	0.05	1
191	16	MET	CB	C	32.81	0.05	1
194	16	MET	N	N	120.48	0.05	1
195	17	LEU	H	H	8.18	0.05	1
196	17	LEU	HA	H	4.05	0.05	1
197	17	LEU	HB2	H	2.07	0.05	1
198	17	LEU	HB3	H	1.55	0.05	1
199	17	LEU	HG	H	1.16	0.05	1
203	17	LEU	CA	C	58.26	0.05	1
204	17	LEU	CB	C	42.33	0.05	1
208	17	LEU	N	N	119.70	0.05	1
209	18	GLU	H	H	8.14	0.05	1
210	18	GLU	HA	H	3.68	0.05	1
211	18	GLU	HB2	H	2.07	0.05	2
212	18	GLU	HB3	H	1.97	0.05	2
216	18	GLU	CA	C	60.06	0.05	1
217	18	GLU	CB	C	29.28	0.05	1
220	18	GLU	N	N	117.32	0.05	1
221	19	LYS	H	H	8.08	0.05	1
222	19	LYS	HA	H	3.83	0.05	1
223	19	LYS	HB2	H	1.93	0.05	2
224	19	LYS	HB3	H	1.89	0.05	2
225	19	LYS	HG2	H	1.47	0.05	2
226	19	LYS	HG3	H	1.34	0.05	2
227	19	LYS	HD2	H	1.67	0.05	2
228	19	LYS	HD3	H	1.65	0.05	2
233	19	LYS	CA	C	58.19	0.05	1
234	19	LYS	CB	C	32.43	0.05	1
238	19	LYS	N	N	120.30	0.05	1
240	20	ASN	H	H	8.48	0.05	1
241	20	ASN	HA	H	4.32	0.05	1
242	20	ASN	HB2	H	2.91	0.05	1

243	20	ASN	HB3	H	2.33	0.05	1
244	20	ASN	HD21	H	7.80	0.05	1
245	20	ASN	HD22	H	6.84	0.05	1
247	20	ASN	CA	C	55.54	0.05	1
248	20	ASN	CB	C	37.29	0.05	1
250	20	ASN	N	N	118.48	0.05	1
251	20	ASN	ND2	N	109.23	0.05	1
252	21	ALA	H	H	8.49	0.05	1
253	21	ALA	HA	H	4.14	0.05	1
254	21	ALA	HB	H	1.57	0.05	1
256	21	ALA	CA	C	55.37	0.05	1
257	21	ALA	CB	C	18.50	0.05	1
258	21	ALA	N	N	122.58	0.05	1
259	22	ARG	H	H	8.85	0.05	1
260	22	ARG	HA	H	3.93	0.05	1
261	22	ARG	HB2	H	1.77	0.05	2
262	22	ARG	HB3	H	1.55	0.05	2
265	22	ARG	HD2	H	0.64	0.05	4
273	22	ARG	CA	C	60.05	0.05	1
274	22	ARG	CB	C	29.75	0.05	1
278	22	ARG	N	N	118.82	0.05	1
282	23	GLU	H	H	8.37	0.05	1
283	23	GLU	HA	H	4.04	0.05	1
284	23	GLU	HB2	H	2.12	0.05	2
285	23	GLU	HB3	H	1.98	0.05	2
286	23	GLU	HG2	H	2.54	0.05	2
287	23	GLU	HG3	H	2.27	0.05	2
289	23	GLU	CA	C	59.52	0.05	1
290	23	GLU	CB	C	29.91	0.05	1
293	23	GLU	N	N	120.08	0.05	1
294	24	ALA	H	H	8.17	0.05	1
295	24	ALA	HA	H	3.72	0.05	1
296	24	ALA	HB	H	1.51	0.05	1
298	24	ALA	CA	C	55.85	0.05	1
299	24	ALA	CB	C	19.39	0.05	1
300	24	ALA	N	N	121.70	0.05	1
301	25	VAL	H	H	8.07	0.05	1
302	25	VAL	HA	H	3.50	0.05	1
303	25	VAL	HB	H	2.39	0.05	1
304	25	VAL	HG1	H	1.06	0.05	2
305	25	VAL	HG2	H	1.07	0.05	2
307	25	VAL	CA	C	66.86	0.05	1
308	25	VAL	CB	C	31.92	0.05	1
311	25	VAL	N	N	114.86	0.05	1
312	26	LYS	H	H	7.79	0.05	1
313	26	LYS	HA	H	4.19	0.05	1
314	26	LYS	HB2	H	2.01	0.05	1
316	26	LYS	HG2	H	1.55	0.05	2
317	26	LYS	HG3	H	1.42	0.05	2
318	26	LYS	HD2	H	1.72	0.05	2
319	26	LYS	HD3	H	1.72	0.05	2
320	26	LYS	HE2	H	2.98	0.05	2
324	26	LYS	CA	C	59.17	0.05	1
325	26	LYS	CB	C	32.39	0.05	1
329	26	LYS	N	N	120.33	0.05	1
331	27	GLU	H	H	8.22	0.05	1
332	27	GLU	HA	H	4.01	0.05	1

333	27	GLU	HB2	H	1.98	0.05	1
335	27	GLU	HG2	H	2.68	0.05	1
336	27	GLU	HG3	H	2.23	0.05	1
338	27	GLU	CA	C	59.28	0.05	1
339	27	GLU	CB	C	29.61	0.05	1
342	27	GLU	N	N	118.86	0.05	1
343	28	LEU	H	H	7.89	0.05	1
344	28	LEU	HA	H	4.31	0.05	1
345	28	LEU	HB2	H	1.82	0.05	2
347	28	LEU	HG	H	1.55	0.05	2
348	28	LEU	HD1	H	0.89	0.05	2
351	28	LEU	CA	C	55.48	0.05	1
352	28	LEU	CB	C	43.76	0.05	1
356	28	LEU	N	N	116.97	0.05	1
357	29	GLY	H	H	7.94	0.05	1
358	29	GLY	HA2	H	3.96	0.05	2
359	29	GLY	HA3	H	3.93	0.05	2
361	29	GLY	CA	C	46.45	0.05	1
362	29	GLY	N	N	108.64	0.05	1
363	30	ILE	H	H	7.23	0.05	1
364	30	ILE	HA	H	4.26	0.05	1
365	30	ILE	HB	H	1.78	0.05	1
366	30	ILE	HG12	H	1.41	0.05	2
367	30	ILE	HG13	H	0.89	0.05	2
368	30	ILE	HG2	H	1.02	0.05	2
371	30	ILE	CA	C	60.24	0.05	1
372	30	ILE	CB	C	40.99	0.05	1
376	30	ILE	N	N	116.07	0.05	1
377	31	ASP	H	H	8.44	0.05	1
378	31	ASP	HA	H	4.79	0.05	1
379	31	ASP	HB2	H	2.67	0.05	2
380	31	ASP	HB3	H	2.56	0.05	2
382	31	ASP	CA	C	53.67	0.05	1
383	31	ASP	CB	C	41.50	0.05	1
385	31	ASP	N	N	124.62	0.05	1
386	32	ALA	H	H	8.07	0.05	1
387	32	ALA	HA	H	5.07	0.05	1
388	32	ALA	HB	H	0.94	0.05	1
390	32	ALA	CA	C	50.69	0.05	1
391	32	ALA	CB	C	23.11	0.05	1
392	32	ALA	N	N	124.68	0.05	1
393	33	GLU	H	H	8.50	0.05	1
394	33	GLU	HA	H	4.61	0.05	1
395	33	GLU	HB2	H	1.83	0.05	1
397	33	GLU	HG2	H	2.23	0.05	1
398	33	GLU	HG3	H	2.07	0.05	1
400	33	GLU	CA	C	54.44	0.05	1
401	33	GLU	CB	C	32.63	0.05	1
404	33	GLU	N	N	120.83	0.05	1
405	34	PHE	H	H	8.85	0.05	1
406	34	PHE	HA	H	5.23	0.05	1
407	34	PHE	HB2	H	2.96	0.05	3
408	34	PHE	HB3	H	2.55	0.05	3
409	34	PHE	HD1	H	7.08	0.05	1
410	34	PHE	HD2	H	7.08	0.05	1
411	34	PHE	HE1	H	7.14	0.05	1
412	34	PHE	HE2	H	7.14	0.05	1

413	34	PHE	HZ	H	6.92	0.05	1
415	34	PHE	CA	C	57.32	0.05	1
416	34	PHE	CB	C	41.95	0.05	1
423	34	PHE	N	N	122.56	0.05	1
424	35	GLU	H	H	9.08	0.05	1
425	35	GLU	HA	H	4.59	0.05	1
426	35	GLU	HB2	H	1.82	0.05	2
427	35	GLU	HB3	H	1.80	0.05	2
428	35	GLU	HG2	H	2.09	0.05	1
429	35	GLU	HG3	H	1.96	0.05	1
431	35	GLU	CA	C	54.82	0.05	1
432	35	GLU	CB	C	33.58	0.05	1
435	35	GLU	N	N	124.20	0.05	1
436	36	LYS	H	H	8.80	0.05	1
437	36	LYS	HA	H	4.83	0.05	1
438	36	LYS	HB2	H	1.67	0.05	2
440	36	LYS	HG2	H	1.33	0.05	2
444	36	LYS	HE2	H	2.87	0.05	1
448	36	LYS	CA	C	55.52	0.05	1
449	36	LYS	CB	C	32.26	0.05	1
453	36	LYS	N	N	125.84	0.05	1
455	37	ILE	H	H	8.54	0.05	1
456	37	ILE	HA	H	4.23	0.05	1
457	37	ILE	HB	H	1.38	0.05	1
458	37	ILE	HG12	H	0.89	0.05	2
461	37	ILE	HD1	H	0.35	0.05	2
463	37	ILE	CA	C	61.26	0.05	1
464	37	ILE	CB	C	37.85	0.05	1
468	37	ILE	N	N	127.75	0.05	1
469	38	LYS	H	H	8.32	0.05	1
470	38	LYS	HA	H	5.19	0.05	1
471	38	LYS	HB2	H	1.77	0.05	1
472	38	LYS	HB3	H	1.74	0.05	1
473	38	LYS	HG2	H	1.26	0.05	2
477	38	LYS	HE2	H	2.91	0.05	2
481	38	LYS	CA	C	55.18	0.05	1
482	38	LYS	CB	C	35.24	0.05	1
486	38	LYS	N	N	123.69	0.05	1
488	39	GLU	H	H	7.48	0.05	1
489	39	GLU	HA	H	3.68	0.05	1
490	39	GLU	HB2	H	1.96	0.05	1
492	39	GLU	HG2	H	2.42	0.05	2
493	39	GLU	HG3	H	2.23	0.05	2
495	39	GLU	CA	C	57.47	0.05	1
496	39	GLU	CB	C	30.77	0.05	1
499	39	GLU	N	N	119.78	0.05	1
501	40	MET	HA	H	4.34	0.05	1
508	40	MET	CA	C	57.47	0.05	1
509	40	MET	CB	C	30.77	0.05	1
513	41	ASP	H	H	9.02	0.05	1
514	41	ASP	HA	H	4.30	0.05	1
515	41	ASP	HB2	H	2.68	0.05	1
516	41	ASP	HB3	H	2.62	0.05	1
518	41	ASP	CA	C	57.47	0.05	1
519	41	ASP	CB	C	39.34	0.05	1
521	41	ASP	N	N	117.05	0.05	1
522	42	GLN	H	H	7.19	0.05	1

523	42	GLN	HA	H	4.13	0.05	1
524	42	GLN	HB2	H	2.12	0.05	2
525	42	GLN	HB3	H	2.11	0.05	2
526	42	GLN	HG2	H	2.44	0.05	2
527	42	GLN	HG3	H	2.41	0.05	2
528	42	GLN	HE21	H	7.74	0.05	1
529	42	GLN	HE22	H	6.88	0.05	1
531	42	GLN	CA	C	58.44	0.05	1
532	42	GLN	CB	C	29.71	0.05	1
535	42	GLN	N	N	119.19	0.05	1
536	42	GLN	NE2	N	112.82	0.05	1
537	43	ILE	H	H	8.07	0.05	1
538	43	ILE	HA	H	3.52	0.05	1
539	43	ILE	HB	H	2.12	0.05	1
540	43	ILE	HG12	H	0.97	0.05	2
541	43	ILE	HG13	H	0.96	0.05	2
543	43	ILE	HD1	H	0.86	0.05	4
545	43	ILE	CA	C	66.27	0.05	1
546	43	ILE	CB	C	38.50	0.05	1
550	43	ILE	N	N	118.18	0.05	1
551	44	LEU	H	H	8.41	0.05	1
552	44	LEU	HA	H	4.09	0.05	1
553	44	LEU	HB2	H	1.80	0.05	1
554	44	LEU	HB3	H	1.47	0.05	1
555	44	LEU	HG	H	0.85	0.05	1
559	44	LEU	CA	C	57.68	0.05	1
560	44	LEU	CB	C	41.31	0.05	1
564	44	LEU	N	N	118.18	0.05	1
565	45	GLU	H	H	7.91	0.05	1
566	45	GLU	HA	H	3.99	0.05	1
567	45	GLU	HB2	H	2.15	0.05	1
568	45	GLU	HB3	H	2.06	0.05	1
569	45	GLU	HG2	H	2.39	0.05	1
570	45	GLU	HG3	H	2.28	0.05	1
572	45	GLU	CA	C	59.01	0.05	1
573	45	GLU	CB	C	29.53	0.05	1
576	45	GLU	N	N	121.46	0.05	1
577	46	ALA	H	H	7.37	0.05	1
578	46	ALA	HA	H	4.25	0.05	1
579	46	ALA	HB	H	1.32	0.05	1
581	46	ALA	CA	C	52.64	0.05	1
582	46	ALA	CB	C	18.98	0.05	1
583	46	ALA	N	N	119.26	0.05	1
584	47	GLY	H	H	7.67	0.05	1
585	47	GLY	HA2	H	4.02	0.05	1
586	47	GLY	HA3	H	3.64	0.05	1
588	47	GLY	CA	C	45.23	0.05	1
589	47	GLY	N	N	104.24	0.05	1
590	48	LEU	H	H	6.91	0.05	1
591	48	LEU	HA	H	4.25	0.05	1
592	48	LEU	HB2	H	1.50	0.05	1
593	48	LEU	HB3	H	1.38	0.05	1
594	48	LEU	HG	H	0.41	0.05	1
598	48	LEU	CA	C	56.80	0.05	1
599	48	LEU	CB	C	41.52	0.05	1
603	48	LEU	N	N	120.48	0.05	1
604	49	THR	H	H	7.72	0.05	1

605	49	THR	HA	H	4.26	0.05	1
606	49	THR	HB	H	4.39	0.05	1
610	49	THR	CA	C	61.03	0.05	1
611	49	THR	CB	C	68.97	0.05	1
613	49	THR	N	N	110.40	0.05	1
614	50	ALA	H	H	7.54	0.05	1
615	50	ALA	HA	H	4.39	0.05	1
616	50	ALA	HB	H	1.32	0.05	1
618	50	ALA	CA	C	51.50	0.05	1
619	50	ALA	CB	C	44.68	0.05	1
620	50	ALA	N	N	123.58	0.05	1
621	51	LEU	H	H	8.17	0.05	1
622	51	LEU	HA	H	4.54	0.05	1
623	51	LEU	HB2	H	1.68	0.05	2
624	51	LEU	HB3	H	1.39	0.05	2
625	51	LEU	HG	H	1.49	0.05	1
626	51	LEU	HD1	H	0.95	0.05	2
630	51	LEU	CB	C	44.68	0.05	1
634	51	LEU	N	N	116.95	0.05	1
635	52	PRO	HA	H	4.89	0.05	1
636	52	PRO	HB2	H	2.30	0.05	1
637	52	PRO	HB3	H	2.05	0.05	1
638	52	PRO	HG2	H	1.98	0.05	1
639	52	PRO	HG3	H	1.91	0.05	1
640	52	PRO	HD2	H	3.87	0.05	2
641	52	PRO	HD3	H	3.70	0.05	2
643	52	PRO	CA	C	62.38	0.05	1
644	52	PRO	CB	C	35.99	0.05	1
648	53	GLY	H	H	9.07	0.05	1
649	53	GLY	HA2	H	5.20	0.05	1
650	53	GLY	HA3	H	3.71	0.05	1
652	53	GLY	CA	C	45.30	0.05	1
653	53	GLY	N	N	103.96	0.05	1
654	54	LEU	H	H	9.11	0.05	1
655	54	LEU	HA	H	5.79	0.05	1
656	54	LEU	HB2	H	1.66	0.05	2
657	54	LEU	HB3	H	1.62	0.05	2
659	54	LEU	HD1	H	0.81	0.05	4
662	54	LEU	CA	C	54.56	0.05	1
663	54	LEU	CB	C	47.03	0.05	1
667	54	LEU	N	N	122.90	0.05	1
668	55	ALA	H	H	9.84	0.05	1
669	55	ALA	HA	H	5.39	0.05	1
670	55	ALA	HB	H	1.52	0.05	1
672	55	ALA	CA	C	50.49	0.05	1
673	55	ALA	CB	C	24.01	0.05	1
674	55	ALA	N	N	131.69	0.05	1
675	56	VAL	H	H	8.31	0.05	1
676	56	VAL	HA	H	4.57	0.05	1
677	56	VAL	HB	H	1.81	0.05	1
678	56	VAL	HG1	H	0.89	0.05	1
679	56	VAL	HG2	H	0.76	0.05	1
681	56	VAL	CA	C	60.13	0.05	1
682	56	VAL	CB	C	34.04	0.05	1
685	56	VAL	N	N	119.12	0.05	1
686	57	ASP	H	H	10.17	0.05	1
687	57	ASP	HA	H	4.44	0.05	1

688	57	ASP	HB2	H	2.99	0.05	1
689	57	ASP	HB3	H	2.66	0.05	1
691	57	ASP	CA	C	56.16	0.05	1
692	57	ASP	CB	C	40.05	0.05	1
694	57	ASP	N	N	130.69	0.05	1
695	58	GLY	H	H	8.98	0.05	1
696	58	GLY	HA2	H	4.20	0.05	2
697	58	GLY	HA3	H	3.54	0.05	2
699	58	GLY	CA	C	45.31	0.05	1
700	58	GLY	N	N	103.37	0.05	1
701	59	GLU	H	H	7.91	0.05	1
702	59	GLU	HA	H	4.61	0.05	1
703	59	GLU	HB2	H	2.08	0.05	1
704	59	GLU	HB3	H	1.98	0.05	1
705	59	GLU	HG2	H	2.41	0.05	1
706	59	GLU	HG3	H	2.28	0.05	1
708	59	GLU	CA	C	54.65	0.05	1
709	59	GLU	CB	C	31.08	0.05	1
712	59	GLU	N	N	121.63	0.05	1
713	60	LEU	H	H	8.96	0.05	1
714	60	LEU	HA	H	4.07	0.05	1
715	60	LEU	HB2	H	1.63	0.05	2
718	60	LEU	HD1	H	0.78	0.05	2
721	60	LEU	CA	C	56.82	0.05	1
722	60	LEU	CB	C	42.15	0.05	1
726	60	LEU	N	N	129.66	0.05	1
727	61	LYS	H	H	9.08	0.05	1
728	61	LYS	HA	H	4.73	0.05	1
729	61	LYS	HB2	H	1.91	0.05	1
730	61	LYS	HB3	H	1.71	0.05	1
739	61	LYS	CA	C	54.90	0.05	1
740	61	LYS	CB	C	35.62	0.05	1
744	61	LYS	N	N	122.44	0.05	1
746	62	ILE	H	H	7.73	0.05	1
747	62	ILE	HA	H	4.35	0.05	1
748	62	ILE	HB	H	1.76	0.05	1
749	62	ILE	HG12	H	1.32	0.05	1
750	62	ILE	HG13	H	0.84	0.05	1
751	62	ILE	HG2	H	1.14	0.05	1
752	62	ILE	HD1	H	0.72	0.05	1
754	62	ILE	CA	C	59.42	0.05	1
755	62	ILE	CB	C	42.21	0.05	1
759	62	ILE	N	N	117.86	0.05	1
760	63	MET	H	H	8.94	0.05	1
761	63	MET	HA	H	4.98	0.05	1
762	63	MET	HB2	H	2.13	0.05	1
763	63	MET	HB3	H	1.94	0.05	1
764	63	MET	HG2	H	2.46	0.05	2
765	63	MET	HG3	H	2.27	0.05	2
768	63	MET	CA	C	55.65	0.05	1
769	63	MET	CB	C	36.01	0.05	1
772	63	MET	N	N	125.24	0.05	1
773	64	GLY	H	H	8.79	0.05	1
774	64	GLY	HA2	H	3.94	0.05	2
775	64	GLY	HA3	H	3.43	0.05	1
777	64	GLY	CA	C	46.38	0.05	1
778	64	GLY	N	N	111.58	0.05	1

779	65	ARG	H	H	7.17	0.05	1
780	65	ARG	HA	H	4.62	0.05	1
781	65	ARG	HB2	H	1.83	0.05	2
782	65	ARG	HB3	H	1.65	0.05	2
783	65	ARG	HG2	H	1.22	0.05	2
784	65	ARG	HG3	H	1.03	0.05	2
785	65	ARG	HD2	H	3.10	0.05	1
793	65	ARG	CA	C	53.88	0.05	1
794	65	ARG	CB	C	32.09	0.05	1
798	65	ARG	N	N	112.57	0.05	1
802	66	VAL	H	H	8.39	0.05	1
803	66	VAL	HA	H	3.74	0.05	1
804	66	VAL	HB	H	1.98	0.05	1
805	66	VAL	HG1	H	0.98	0.05	1
806	66	VAL	HG2	H	0.73	0.05	1
808	66	VAL	CA	C	62.22	0.05	1
809	66	VAL	CB	C	31.50	0.05	1
812	66	VAL	N	N	117.34	0.05	1
813	67	ALA	H	H	6.01	0.05	1
814	67	ALA	HA	H	4.49	0.05	1
815	67	ALA	HB	H	1.20	0.05	1
817	67	ALA	CA	C	50.74	0.05	1
818	67	ALA	CB	C	21.41	0.05	1
819	67	ALA	N	N	128.70	0.05	1
820	68	SER	H	H	8.82	0.05	1
821	68	SER	HA	H	4.29	0.05	1
822	68	SER	HB2	H	4.42	0.05	1
826	68	SER	CA	C	57.27	0.05	1
827	68	SER	CB	C	65.30	0.05	1
828	68	SER	N	N	117.13	0.05	1
829	69	LYS	H	H	9.12	0.05	1
830	69	LYS	HA	H	3.75	0.05	1
833	69	LYS	HG2	H	1.62	0.05	2
841	69	LYS	CA	C	60.63	0.05	1
842	69	LYS	CB	C	31.54	0.05	1
846	69	LYS	N	N	121.03	0.05	1
848	70	GLU	H	H	8.72	0.05	1
849	70	GLU	HA	H	3.82	0.05	1
850	70	GLU	HB2	H	1.97	0.05	1
852	70	GLU	HG2	H	2.42	0.05	1
853	70	GLU	HG3	H	2.29	0.05	1
855	70	GLU	CA	C	60.40	0.05	1
856	70	GLU	CB	C	29.03	0.05	1
859	70	GLU	N	N	117.40	0.05	1
860	71	GLU	H	H	8.01	0.05	1
861	71	GLU	HA	H	3.93	0.05	1
862	71	GLU	HB2	H	1.97	0.05	2
863	71	GLU	HB3	H	1.96	0.05	2
864	71	GLU	HG2	H	2.41	0.05	1
865	71	GLU	HG3	H	2.18	0.05	1
867	71	GLU	CA	C	59.49	0.05	1
868	71	GLU	CB	C	29.92	0.05	1
871	71	GLU	N	N	121.14	0.05	1
872	72	ILE	H	H	8.19	0.05	1
873	72	ILE	HA	H	3.50	0.05	1
874	72	ILE	HB	H	1.68	0.05	1
877	72	ILE	HG2	H	0.75	0.05	2

878	72	ILE	HD1	H	0.51	0.05	2
880	72	ILE	CA	C	65.39	0.05	1
881	72	ILE	CB	C	37.81	0.05	1
885	72	ILE	N	N	119.50	0.05	1
886	73	LYS	H	H	8.61	0.05	1
887	73	LYS	HA	H	3.73	0.05	1
888	73	LYS	HB2	H	1.88	0.05	1
889	73	LYS	HB3	H	1.88	0.05	1
890	73	LYS	HG2	H	1.23	0.05	2
898	73	LYS	CA	C	60.96	0.05	1
899	73	LYS	CB	C	32.25	0.05	1
903	73	LYS	N	N	118.99	0.05	1
905	74	LYS	H	H	7.36	0.05	1
906	74	LYS	HA	H	4.03	0.05	1
907	74	LYS	HB2	H	1.96	0.05	1
908	74	LYS	HB3	H	1.68	0.05	1
909	74	LYS	HG2	H	1.44	0.05	1
911	74	LYS	HD2	H	1.57	0.05	1
913	74	LYS	HE2	H	2.95	0.05	1
917	74	LYS	CA	C	59.14	0.05	1
918	74	LYS	CB	C	32.10	0.05	1
922	74	LYS	N	N	117.43	0.05	1
924	75	ILE	H	H	7.66	0.05	1
925	75	ILE	HA	H	3.92	0.05	1
926	75	ILE	HB	H	2.07	0.05	1
927	75	ILE	HG12	H	1.23	0.05	1
928	75	ILE	HG13	H	0.75	0.05	1
929	75	ILE	HG2	H	0.87	0.05	1
932	75	ILE	CA	C	63.94	0.05	1
933	75	ILE	CB	C	38.45	0.05	1
937	75	ILE	N	N	118.82	0.05	1
938	76	LEU	H	H	7.54	0.05	1
939	76	LEU	HA	H	4.24	0.05	1
940	76	LEU	HB2	H	1.86	0.05	2
941	76	LEU	HB3	H	1.84	0.05	2
942	76	LEU	HG	H	1.57	0.05	1
943	76	LEU	HD1	H	0.74	0.05	2
946	76	LEU	CA	C	55.05	0.05	1
947	76	LEU	CB	C	42.49	0.05	1
951	76	LEU	N	N	118.18	0.05	1
952	77	SER	H	H	7.24	0.05	1
953	77	SER	HA	H	4.29	0.05	1
954	77	SER	HB2	H	3.89	0.05	2
958	77	SER	CA	C	60.44	0.05	1
959	77	SER	CB	C	65.27	0.05	1
960	77	SER	N	N	119.27	0.05	1